Routescore: Punching the Ticket to More Efficient Materials Development

Martin Seifrid[a,b], Riley J. Hickman[a,b], Andrés Aguilar-Granda[a,b], Cyrille Lavigne[b], Jenya Vestfrid[a,b], Tony C. Wu[a,b], Théophile Gaudin[b,c], Emily J. Hopkins[a], Alán Aspuru-Guzik[a,b,d,e]

[a]*Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada*
[b]*Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada*
[c]*IBM Research Zürich, 8803 Rüschlikon, Zürich, Switzerland*
[d]*CIFAR Artificial Intelligence Research Chair, Vector Institute, Toronto, ON M5S 1M1, Canada*
[e]*Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, ON M5S 1M1, Canada*

ABSTRACT: Self-driving labs, in the form of automated experimentation platforms guided by machine learning algorithms have emerged as a potential solution to the need for accelerated science. While new tools for automated analysis and characterization are being developed at a steady rate, automated synthesis remains the bottleneck in the chemical space accessible to self-driving labs. Combining automated and manual synthesis efforts immediately significantly expands the explorable chemical space. To effectively direct the different capabilities of automated (higher throughput and less labor) and manual synthesis (greater chemical versatility), we describe a protocol, the RouteScore, that quantifies the cost of combined synthetic routes. In this work, the RouteScore is used to determine the most efficient synthetic route to a well-known pharmaceutical (structure-oriented optimization), and to simulate a self-driving lab that finds the most easily synthesizable organic laser molecule with specific photophysical properties from a space of ~3500 possible molecules (property-oriented optimization). These two examples demonstrate the power and generality of our approach in mixed synthetic planning and optimization.

Introduction

Molecular design and discovery is a universal challenge across the chemical sciences, which requires exploring a vast chemical space.[1–3] Self-driving labs, also known as materials acceleration platforms (MAPs), have the potential to make faster, more efficient progress by "closing" the chemical discovery loop: integrating property prediction, synthesis, analysis, characterization and experiment planning.[4–6] One of the key challenges in building self-driving labs is developing a platform capable of autonomously performing all experiments from synthesis to characterization. Automated synthesis platforms (ASPs) are therefore an integral element of MAPs: synthesis is the engine that drives exploration of chemical space. At the moment, ASPs are only capable of performing a very limited set of reactions compared to human chemists.[7–12] As a result, the chemical space accessible to MAPs is limited by the reactions the ASP can perform, as well as the price and availability of the starting material library since high-throughput experiments often require more material than manual synthesis. Consequently, molecules incorporating starting materials that are unavailable or cost-prohibitive cannot be explored, even though computations may predict them to have highly desirable properties. To this end, we envision a combined synthetic strategy including both manual and automated synthesis (Figure 1), where human chemists synthesize the molecules inaccessible to the ASP while taking advantage of its increased throughput to more rapidly travel through chemical space.

The combined automated and manual synthetic approach to traversing chemical space can be likened to a subway system in a large city. In this "chemical metropolis," the cost of the starting materials is analogous to rental or housing prices: the closer you are to your target, the more expensive the starting materials. In this analogy, the subway lines – fast and efficient with limited stops – are the reactions carried out by the ASP. Manual reactions – slow and costly, but much more versatile – are walking to or from the subway station. Finally, the "fare" for traveling through chemical space are the monetary, material and time costs of carrying out the syntheses.

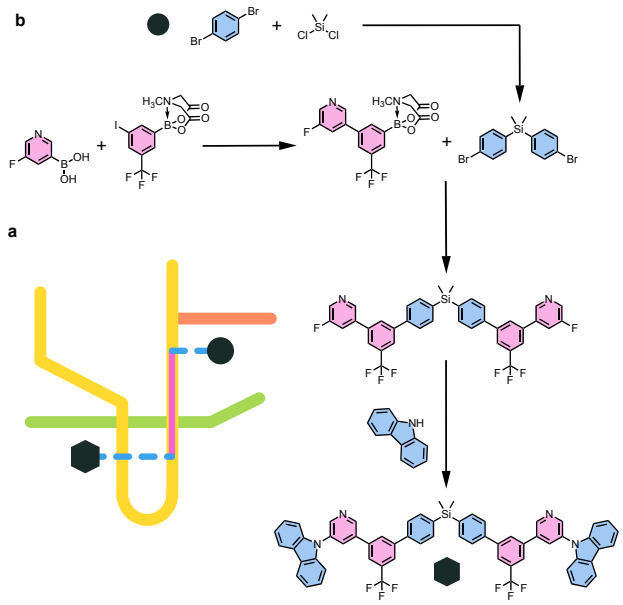

Figure 1. Subway map of chemical space (a) depicting travel through chemical space from (b) 1,4-dibromobenzene and dimethyldichlorosilane (circle) to the target molecule (hexagon) using both manual reactions (blue) and automated iterative Suzuki-Miyaura cross-coupling reactions (pink).

Quantifying the difficulty of synthesizing a target molecule is a very important challenge in both synthetic chemistry and cheminformatics. Commonly, synthetic accessibility is quantified based on a variety of structural features of the target molecule, including the number of rings and stereocenters, the complexity of the target molecule's graph representation, or similarity to the starting materials.[13–15] Other approaches consider these factors, as well as more practical considerations, such as the probability of finding a similar molecule or substructure in a database of purchasable starting materials, or the costs of starting materials.[16,17] However, some of these metrics rely on weights for each factor that are assigned based on fitting to expert opinion. In addition to the significant human labor required for determining the weights, this restricts the metric to evaluating only molecules similar to those which were scored by experts. Machine learning (ML)-based approaches for calculating synthetic accessibility have recently been shown to accurately estimate the complexity of a target molecule and synthetic route.[18–21] However, these also face similar limitations in terms of transferability and training. Currently, no synthetic accessibility metrics exist for combined manual and automated synthetic routes.

In this work, we present a new method to evaluate the cost of synthetic routes. The *RouteScore* requires no pre-training or fitting, and is based on objective inputs and weights such as cost of labor and materials and human or robot time. Although it is designed with the "subway map" approach of combined manual and automated synthesis in mind, the *RouteScore* is equally adaptable to fully-automated or fully-manual synthesis. Furthermore, it can be used in both a priori synthetic route planning, as well as in a posteriori evaluation of syntheses. First, we describe how the *RouteScore* can

be used to determine the most efficient synthetic route from many (structure-oriented) by comparing ten different syntheses of a molecule with many known routes, modafinil. Then, we show how, by traveling through the chemical subway map, multi-objective optimization using the *RouteScore* as one of the objectives can be used to determine promising candidate molecules for organic laser molecules (property-oriented).

Results and discussion

The cost of a synthetic route

To select the best synthetic route, we calculate the route's cost per amount of target molecule produced. There are three principal considerations when determining the cost of a reaction: time, money and mass efficiency. The latter two are rarely considered in academic settings.[17] Here, we define the mass cost as the total mass of material required for a reaction, and the monetary cost as the sum of the cost of human and robotic labor and the total cost of the starting materials used in the reaction. This is the factor that will be most variable between laboratories, institutions and countries due to differences in labor and material costs. For the purpose of clarity, we have included a full breakdown of our calculations of the labor costs in Tables S1 and S2. The monetary cost of starting materials synthesized in a previous step along the route is not factored into the monetary cost of a subsequent reaction so as to avoid double-counting. A "step" is defined as a reaction that requires setting up and later cleaning labware, that is to say that one-pot multistep reactions – although they involve multiple chemical transformations – only count for a single step in the *RouteScore*, as these are generally more efficient. Solvents used in the reaction, as well as work-up, purification and washing are not directly factored into the *StepScore* because they usually reflect only a small portion of the cost of a reaction. An easier approach to accounting for such "hidden" costs is to include them in the labor cost. In the case of a priori estimation of the *RouteScore*, the yield should be assumed to be 1. However, estimates of a reaction's yield could also be provided by forward reaction prediction algorithms.[22–24]

We define the total time cost (*TTC*) of combined human and robotic syntheses as follows:

$$TTC = \sqrt{\left(\frac{C_H}{C_M} t_H\right)^2 + \left(\frac{C_M}{C_H} t_M\right)^2} \quad (1)$$

The surface of all possible time costs is a cone with minimum of 0 at $t_H = 0$ and $t_M = 0$ (Figure 2). This results in a linear increase in the *TTC* for any combination of t_H and t_M. In the case where the hourly costs of human (C_H) and machine (C_M) labor are different, the surface is an elliptic cone where the semimajor and semiminor axes correspond to the ratio of C_H and C_M. We generally expect human time to be more expensive. This means that for equal increases in t_H and t_M, an increase in t_H results in an increase of the *TTC* that is proportional to C_H/C_M. Therefore the *RouteScore* will disincentivize reactions that require large t_H. Only taking into account t_H could lead the *RouteScore* to favor reactions that require very large t_M, which is also undesirable.

Based on these considerations, we define the cost of a reaction step along the synthetic route (*StepScore*) to be:

$$StepScore = \sqrt{\left(\frac{C_H}{C_M} t_H\right)^2 + \left(\frac{C_M}{C_H} t_M\right)^2} \times \left(\sum_{H,M} t_{H,M} C_{H,M} + \sum_i n_i C_i\right) \times \sum_i n_i MW_i \quad (2)$$

where n_i is the molar quantity of a given material, C_i is its cost and MW_i is its molecular weight. For purely manual synthesis, C_H, C_M and t_M can be dropped, giving $TTC = t_H$. When determining the *TTC* for automated synthesis, it is also important to account for the human time required for maintenance of the robotic chemist.

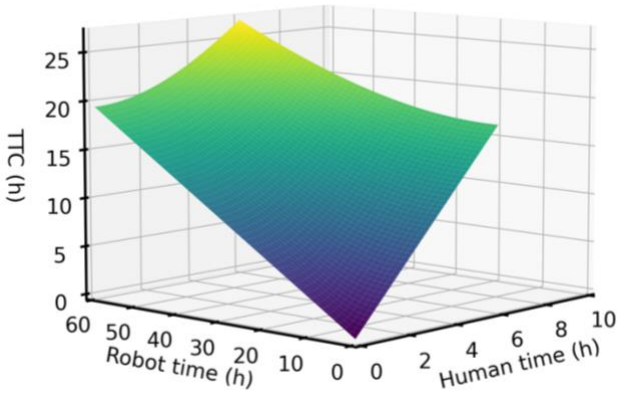

Figure 2. Plot of the TTC as a function of human and robot time.

To make syntheses at different scales comparable, the sum of all *StepScores* is normalized by the quantity of target material produced (n_{Target}). The *RouteScore*, with units of h·$·g·(mol of target molecule)$^{-1}$, can therefore be expressed with this equation:

$$RouteScore = \frac{\sum_N StepScore_N}{n_{Target}} \quad (3)$$

Synthetic route optimization for a well-studied drug molecule

It can be difficult to quantify the efficiency of a diverse set of synthetic routes. To demonstrate the usefulness of the *RouteScore* for addressing this challenge, we selected a drug, modafinil, which has many known synthetic routes (Table S3).[25–32] For each route (Figure 3), we determined the required human time based on our own estimates (see SI for details) and calculated the *RouteScore*. The synthetic routes vary from a patented industrial-scale preparation[31] to a milligram-scale synthesis performed to screen modafinil's anti-inflammatory activity.[27]

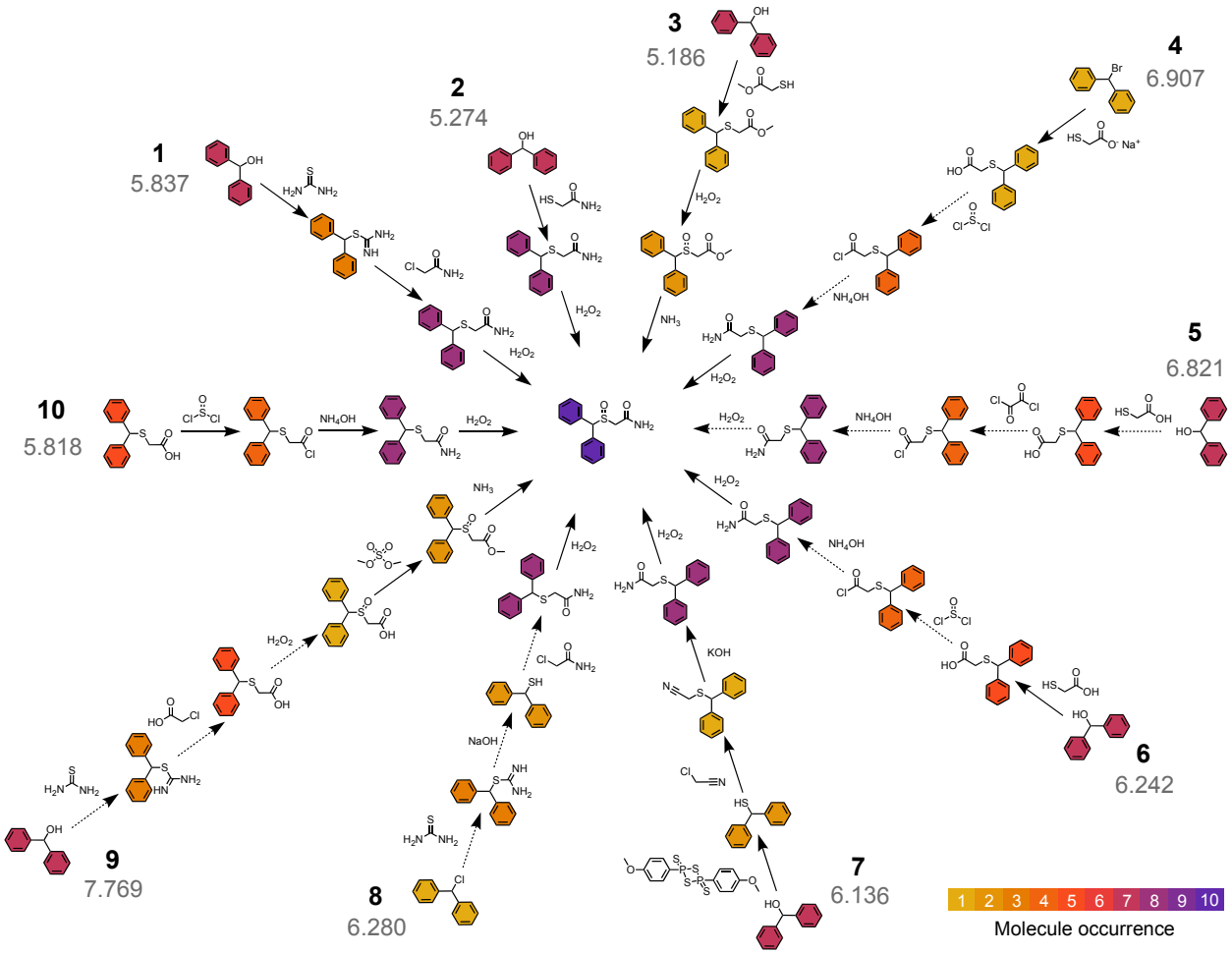

Figure 3. Ten routes to synthesize modafinil, with route number bolded and log(*RouteScore*) in grey underneath. The rings of each molecule are colored based on how often that molecule appears in the 10 routes. Dashed arrows represent one-pot multistep reactions, which we treated as a single step.

We find that the scale of the synthetic route and the number of steps do not correlate strongly with the *RouteScore* (Figure 4). Routes 1 and 3 both start from diphenyl methanol, take three steps and have similar overall yields (65-66%, Figure S3). The main difference between the two routes is in the procedure. Repeated drying and purification by recrystallization are labor-intensive (Table S4), and often involve loss of 5-10% of the product. Route 1 requires numerous recrystallizations, which raises the total labor time from 4 hours (route 3) to 6.5 hours (route 1). Unlike the other routes, route 5 is carried out as a one-pot multistep synthesis in bespoke 3D-printed reactionware, which is intended to minimize human labor required to carry out syntheses. However, this route requires many small operations (e.g., preparing syringes, transferring solutions from one reactor module to another) which add up to 6 hours of labor time. As a result, route 6, which is almost identical to route 5, requires slightly less time (5.5 hours). The routes that include formation of the 2-(benzhydrylthio)acetyl chloride intermediate (routes 4, 5, 6 and 10) are much less efficient, likely due to the extra precautions and labor required to use reagents such as thionyl chloride and oxalyl chloride. Route 7 takes 4 steps, but ends up being less costly (log(*RS*) = 6.136) than routes 5 and 6 despite substantial labor costs (9.25 hours) and a mediocre overall yield (34%) because the monetary cost of each step is quite low (on average $123 per step). Finally, routes 2 and 3, which use Nafion as a catalyst, are the most efficient because they require very little labor, are cheap to carry out and efficiently utilize the catalyst and starting materials to build up the target molecule (Figure

S3). Notably, route 3 is less costly than route 2 despite requiring more labor and having one more step because it uses a much cheaper method of introducing the thioether and amide groups. The 2-mercaptoacetamide reactant costs $2303 CAD/mol, while methyl thioglycolate only costs $29 CAD/mol and the amide can easily be synthesized from ammonia ($83 CAD/mol) at the last step. The effectiveness of this strategy is supported by a similar approach in the patented industrial synthesis (route 9).[31] Although route 9 is carried out on an industrial scale, it is the least efficient (log(*RS*) = 7.700) because it suffers from a below average overall yield of 23% (Figure S3) and requires a significant amount of human labor (9.5 hours). Since the *RouteScore* has identified this industrial-scale synthesis as being inefficient, its quantitative information can be used to translate the advantages of other syntheses of modafinil to a more efficient method to potentially produce large quantities of the target molecule.

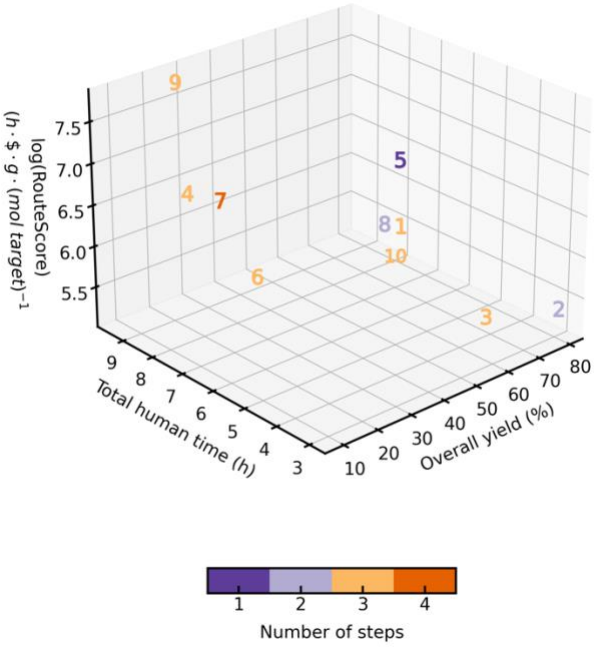

Figure 4. Results of evaluating the 10 modafinil synthetic routes using the *RouteScore*. The horizontal axes correspond to the total human time required to perform each synthesis and the overall yield of the route. The vertical axis corresponds to log(*RouteScore*). Each point is colored based on the number of synthetic steps, and labeled by its route number.

Multi-objective optimization of organic laser molecules

To demonstrate the usefulness of the *RouteScore* approach for searching chemical space, we perform an in silico optimization of optoelectronic properties of potential organic laser molecules.[33] Organic laser molecules in the solid state could be a very interesting technology for portable devices, and is a logical extension of organic light emitting diode technology. The initial set of molecules are those that can be synthesized by two steps of automated iterative Suzuki-Miyaura cross-coupling (iSMC) reactions[7,34] (Figure 5a) from three groups of building blocks – A, B and C – (Figure S4) to form A-B-C-B-A pentamers. The terms "building block" and "fragment" are sometimes used interchangeably in settings that include both computational and synthetic material design, which can cause confusion. Here, the term "building block" refers to a molecule, which has reactive functional groups, that is used as a reactant in the synthetic route. On the other hand, "fragment" refers to a structural template that is used in computational screening. We randomly picked 10 "A" blocks, 11

"B" blocks and 18 "C" blocks from a list of aromatic compounds, resulting in a space of 1980 symmetric pentamers that could be synthesized in an automated fashion. Most of the blocks are commercially available, however three are not (blue in Figure S4c). In our model system, those were prepared by manual synthesis (Figure S5) using procedures in the literature.[35–37] We estimated the human time required for each synthesis based on prior experience (Table S5). Due to compatible functional groups, certain pentamers can be expanded with post-automation manual synthetic steps involving either nucleophilic aromatic substitution[38] (S_NAr) by a carbazole (Figure 5b), or a Buchwald-Hartwig amination[39] (BHA) with 2-bromopyrazine, via a *t*-butoxycarbonyl (Boc) deprotection step (Figure 5c).

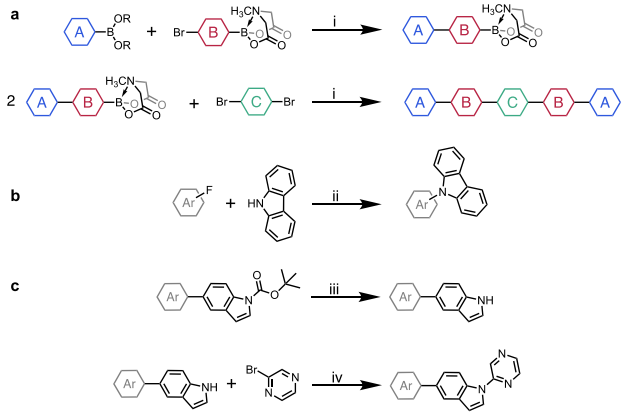

Figure 5. The three syntheses used in our example: iterative Suzuki-Miyaura cross-coupling (a), nucleophilic aromatic substitution (b), and Buchwald-Hartwig amination (c). The following reagents were used for each general type of reacton: (i) XPhos Pd G2, K_3PO_4; (ii) Cs_2CO_3; (iii) K_2CO_3; (iv) $Pd_2(dba)_3$, DavePhos, NaO*t*-Bu. Structures of the organic reagents are provided in Figure S6.

The manually synthesized "C" blocks along with the S_NAr and BHA reactions allow us to explore how adding manual synthetic steps into otherwise automated synthetic exploration of chemical space affects the *RouteScore*. There are 198 pentamers only subjected to manual synthetic modification via Buchwald-Hartwig amination, and 1231 pentamers only modified by manual S_NAr reactions. Finally, there are 49 pentamers that undergo both S_NAr and BHA. For these, we compare the cost of performing either the S_NAr or BHA reactions first. Using only three general types of reactions and 41 total building blocks, we are able to access a chemical space of 3458 molecules.

As expected, we find that the most efficient synthetic routes do not involve any manual synthetic steps after the automated pentamer synthesis ($\log_{10}(\overline{\imath SMC\ auto.}) = 4.94$, Figure 6). The relative cost for manual synthesis of starting materials depends strongly on the particular intermediates and the reactions being carried out. In the *iSMC* set, synthetic routes with the three manually-synthesized "C" blocks ($\log_{10}(\overline{\imath SMC\ man.}) = 7.21$) are ~186 times more costly on average than pentamers synthesized exclusively from commercially available starting materials. Candidate molecules can also be synthesized using S_NAr ($\log_{10}(\overline{S_NAr}) = 7.49$) or BHA ($\log_{10}(\overline{BHA}) = 7.41$) reactions. We find that for the set of 49 molecules that undergo both the S_NAr and BHA reactions, it is less efficient to perform the BHA as the second step ($\log_{10}(\overline{S-B}) = 7.72$), than as the first step $\log_{10}(\overline{B-S}) = 7.69$). The difference in *RouteScore* between the S_NAr-followed-by-BHA (S-B) and BHA-followed-by-S_NAr (B-S) routes is due to the difference in mass of required starting materials for the Boc-deprotection and BHA reactions (Figures S7 and S8). Since the carbazole groups have already been installed at the time of the Boc-deprotection and BHA reactions in the S-B routes, the mass of starting material required to produce the same quantity (mols) of target molecule is greater than for B-S routes. As a result, the *RouteScore* of S-B routes is ~4% greater than that of B-S routes.

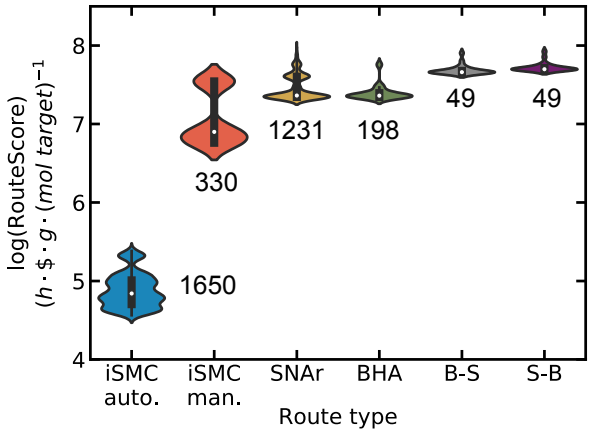

Figure 6. Violin plots of log(*RouteScore*) based on the type of synthesis used in the route. The numbers next to each violin correspond to the number of molecules in each set. The abbreviations are as follows: *iSMC auto.*, molecules synthesized only by automated iSMC; *iSMC man.*, molecules synthesized only by automated iSMC and manual building block synthesis; *SNAr*, molecules involving post-functionalization with only S_NAr reactions; *BHA*, molecules involving post-functionalization with only BHA reactions; *B-S*, molecules where BHA reactions were performed before S_NAr; *S-B*, molecules where S_NAr reactions were performed before BHA. The white dot represents the median value, and the black box indicates the interquartile range.

One of the primary goals of MAPs is achieving efficient inverse design of functional molecules.[5,40] Rather than enumerating large combinatorial spaces of molecules with potentially costly property measurements, the inverse design paradigm seeks to discover molecules starting from a desired property or set of properties. Selecting molecules that satisfy multiple predefined targets simultaneously (e.g. strong emission in a particular wavelength range, low synthetic cost) is a critical but challenging decision-making process, especially when the property measurements are time- or resource-intensive. In this section, we simulate a MAP for the inverse design of organic laser molecules.

The computationally predicted properties of laser molecules are optimized using a multi-objective, categorical variable approach. As objectives, we chose three figures of merit that are important for developing new organic laser molecules[33,41] and the *RouteScore* as objectives for the recently reported deep categorical Bayesian optimizer Gryffin.[42] The four targeted figures of merit in descending order of importance are: (i) maximal fluorescence within a particular spectral range (400-460 nm in this case), (ii) minimal *RouteScore*, (iii) minimal spectral overlap between fluorescence and absorption spectra, and (iv) maximal fluorescence rate. First, maximizing fluorescence within a particular spectral range is necessary for developing a laser of a desired color, arguably the most critical property of any laser device. The *RouteScore* is chosen as the second most important figure of merit to reflect the necessity of finding organic laser molecules that can be synthesized in a cheap and efficient manner. Third, minimizing the spectral overlap corresponds to reducing losses from the self-absorption of emitted light, the inner filter effect.[43] Finally, maximizing the fluorescence rate should improve the quantum efficiency of the laser. The *RouteScore* is calculated as described above, while the other three figures of merit are derived from the results of high-throughput quantum chemical calculations (see SI for details). There are three categorical variables, corresponding to the "A" , "B", and "C" fragments (Figure S1), with 14, 13 and 19 options respectively. This space corresponds to 3458 unique molecules. We use the scalarizing function Chimera[44] to simultaneously optimize the four objectives. Chimera attempts to optimize each objective in order of importance to bring its value within a desired threshold, as described in Ref [44]. We set absolute

tolerances such that roughly 1% of the entire molecular space (34 out of 3458 molecules) satisfies all 4 tolerances simultaneously (Figure 7a). We execute 50 independently seeded optimization runs, each evaluating properties for 500 molecules. Nearing 500 evaluations, we observe asymptotic behavior of the optimizer for each target property. Optimization traces for the four target properties are presented as blue traces in Figure 7b-e.

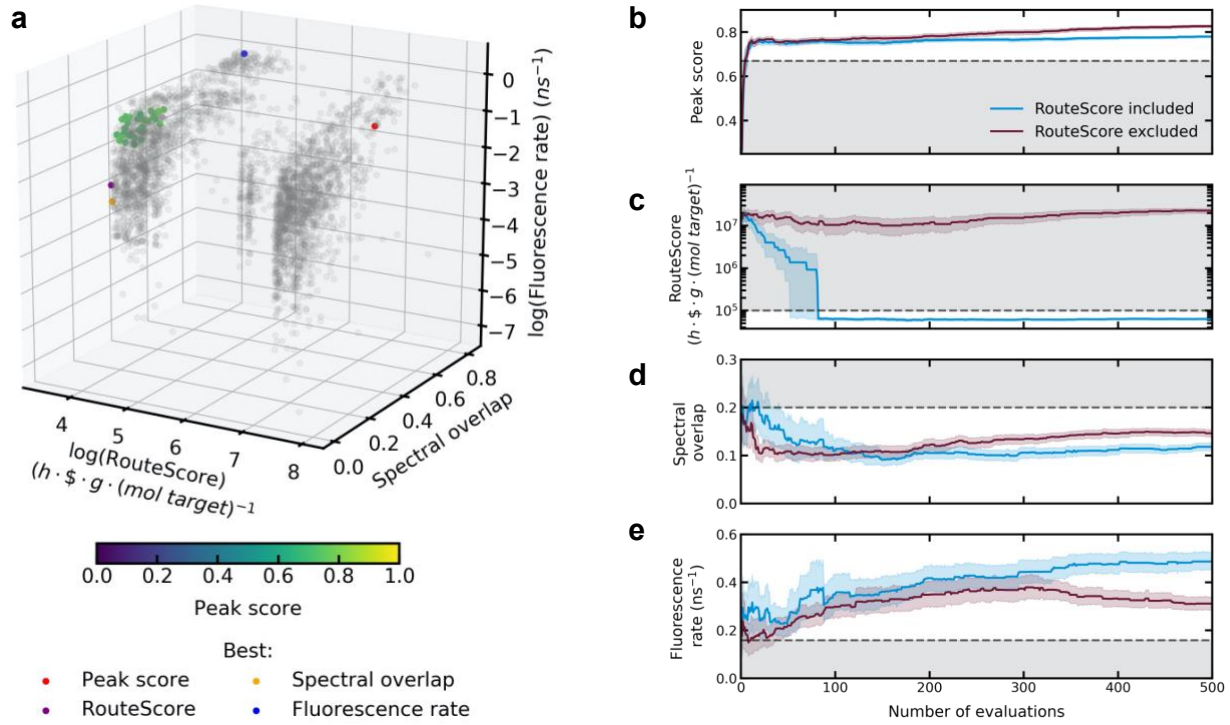

Figure 7. The molecular space for the multi-objective optimization is represented in 4 dimensions (a). The grey points don't satisfy the optimization thresholds. The red, purple, orange and blue points correspond to the molecules with the best peak score, *RouteScore*, spectral overlap and fluorescence rate, respectively (Figure S9). The peak scores of the full molecular space are shown in Figure S10. At each iteration of the multi-objective optimizations using Gryffin and Chimera, we plot the four properties that correspond to the measurement with the best merit: peak score (b), *RouteScore* (c), spectral overlap (d), and fluorescence rate (e). The shaded areas around the curves correspond to the bootstrapped 95% confidence interval. The grey shaded area indicates regions in which tolerances are not satisfied. The dashed lines correspond to the absolute tolerance that must be satisfied for the peak score (> 0.67), *RouteScore* (< 10^5 h·$·g·(mol target molecule)$^{-1}$), spectral overlap (< 0.2) and fluorescence rate (> 0.16 ns^{-1}). All four objectives are optimized simultaneously in the blue traces, while the *RouteScore* is excluded from the set of objectives in the maroon traces.

In this work, we compute the four objective values for all 3458 molecules in our search space before commencing the optimization experiments. As such, we can apply the scalarizing function to the entire dataset a priori and rank the candidate molecules based on the merit returned by Chimera. The 34 satisfactory molecules ordered by the merit-based function constructed from the four objective hierarchy and absolute tolerances are shown in Figure S11, with their objective values in Table S6. Optimizations of the merit-based function should then converge upon the top ranked molecule. In the case of inverse design using Chimera with experimental data, this would not be feasible. Here, we calculate the merit for all the molecules in the search space to evaluate the performance of Gryffin and Chimera.

Gryffin rapidly identifies molecules with fluorescence spectra overlapping significantly with our target region (peak score > 0.67). After achieving the first objective, the *RouteScore* is decreased until its tolerance is satisfied after roughly 80 evaluations while the primary objective remains satisfied. In other words, the algorithm begins to evaluate molecules which have less costly syntheses whose fluorescence spectra fall into the target energy interval in the very first steps of the optimization. The tertiary objective tolerance is satisfied almost immediately after beginning the optimization. As we improve upon the quaternary fluorescence rate objective, we observe a slight regression upon the tertiary objective, i.e. increase in the spectral overlap. To emphasize the effect of including the *RouteScore* in the set of objectives, we conduct additional optimization runs using only three objectives: peak score, spectral overlap, and fluorescence rate. The top 20 molecules according to the merit-based function constructed from this three objective hierarchy and absolute tolerances are shown in Figure S12, with their objective values in Table S7. Optimization traces for these experiments are shown in Figure 7b-e in maroon. Without the additional task of minimizing the *RouteScore,* Gryffin identifies molecules as being meritorious based solely on the properties derived from quantum chemical calculations. As such, molecules identified after 500 iterations have comparable properties to the molecules in the blue traces but are significantly more costly to synthesize (average *RouteScore* > 10^7) in terms of a combination of effort, price and materials needed.

Recently, several studies have highlighted the efficiency of ML-driven experiment planners for achieving inverse design.[45–51] We follow suit for our simulated MAP by quantitatively comparing its aptitude for identifying synthetically feasible laser molecules to that of a simple random sampling strategy. Here we consider the following question: what fraction of total satisfactory molecules can each strategy identify given a budget of 500 evaluations (Figure S14)? In this context, satisfactory refers to a molecule whose properties simultaneously satisfy all the tolerances. The Gryffin + Chimera strategy identifies on average 35±3% of all satisfactory molecules after 500 evaluations, while random sampling identifies only 15±1% of satisfactory candidates (Figure S14). This corresponds to on average about 12 hits with Gryffin + Chimera, but only 5 hits with random sampling. For the entirety of the optimization experiment, the Gryffin + Chimera strategy evaluates on average a greater fraction of total satisfactory molecules, indicating that ML-driven experiment planning strategies yield greater exposure to promising candidates given budgeted resources than does random sampling.

The results of our MAP simulation indicate that the *RouteScore* can be seamlessly used alongside photophysical figures of merit in the multi-objective inverse design of organic laser molecules. In our 50 optimizations, the Gryffin + Chimera strategy identified 12 distinct molecules (Figure 8a), all of which can be synthesized using only automated iSMC reactions. The optimizations overwhelmingly (42% of the time, Figure 8b) identify molecule **1** as the top candidate for synthesis. In contrast, the optimizations that only consider the peak score, spectral overlap and fluorescence rate identify 10 molecules (Figure S15). Molecules **1**, **2** and **3** in the four-objective (with *RouteScore*) optimizations are the same as molecules **F**, **I** and **J** in the three-objective (without *RouteScore*) optimizations. However, these three molecules are only identified as the top choice in 12% of the three-objective optimizations, while they are identified as the top choice in 72% of the four-objective optimizations. Notably, all the molecules identified in the *RouteScore* optimizations contain unusual substitution patterns for organic laser molecules.[33] For example, many of the top molecules are severely sterically hindered due to, e.g., 2,3- or ortho-substitution. This may be related to biases within the choice of building blocks and the target spectral range since 400-460 nm corresponding to relatively high energy violet light. Nonetheless, this design motif may be worth exploring further experimentally and computationally.

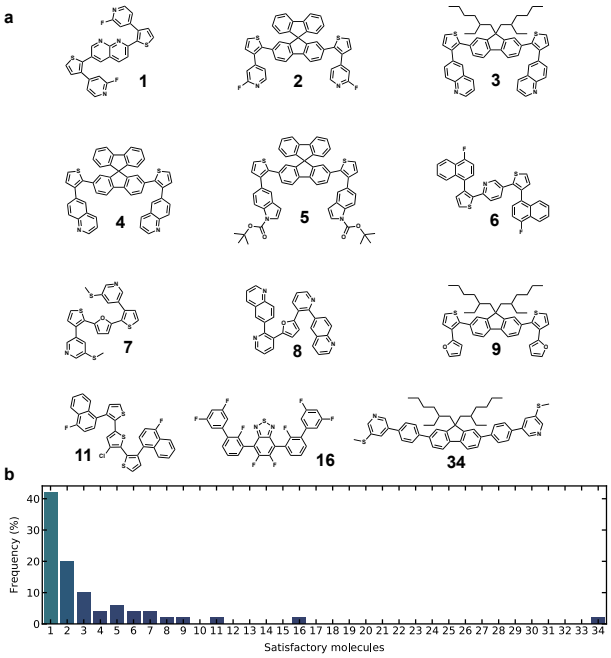

Figure 8. Structures (a) of the top molecules from the four-objective optimizations, numbered by their ranking by merit, and (b) the frequency with which they were found. Molecular structures of all satisfactory molecules are provided in Figure S11.

Conclusion

We have demonstrated a general new approach to quantifying the cost of synthesizing organic molecules, the *RouteScore*, based on factors including the labor and monetary cost of the route, as well as the mass of material consumed. The *RouteScore* promotes more practical considerations about the amount of work required, rather than the elegance of the synthetic route. We have shown how this can be used to select the most efficient synthetic route to a well-known API with numerous reported syntheses. Furthermore, our approach – which takes into account the labor of both manual and automated synthesis – can be used as a tool in self-driving labs to expand the chemical space accessible by MAPs. To demonstrate this principle, we have carried out a multi-objective optimization to select a candidate organic laser molecule based on its fluorescence within a desired wavelength range, its *RouteScore*, the overlap between its absorption and fluorescence spectra and its fluorescence rate using the Gryffin and Chimera algorithms. The ML-driven optimizations efficiently identify top candidate molecules. In addition, optimizations that ignore the *RouteScore* identify molecules with similar predicted photophysical properties, but that are more costly to synthesize. Although we focused on organic materials, this method can be expanded to, e.g., inorganic materials synthesis. In general, the *RouteScore* and subway approaches may be a solution to the challenges of developing self-driving labs.

Although the *RouteScore* is generally robust, there are some important caveats. For example, determining the labor – and its cost – needed for each reaction will require more careful accounting than is typically carried out in academic laboratories. However, we believe that better understanding the underlying costs of material design will have significant benefits. Additionally, it may be desirable to remove the variance between *RouteScore* values in different currencies by normalizing price with respect to some commonly used chemical, similar to the Big Mac index.[52] Further refinement of the *RouteScore* code to calculate the required labor from a chemical descriptive language[9] will reduce the

need for estimation by experts. In addition, although the *RouteScore* can only be most easily compared between laboratories where the material and labor costs are relatively similar, the code released with this work is sufficiently flexible and easy to implement that we hope calculating the *RouteScore* for different laboratories does not impede its adoption.

Supporting Information

The Supplementary Information is available free of charge at [link to SI]:

- Methods
- Supplementary text
- Figure S1 to S15
- Tables S1 to S7

The code and supporting data are freely available as a GitHub repository at https://github.com/aspuru-guzik-group/routescore, and at Zenodo (https://doi.org/10.5281/zenodo.5106659).[53]

Acknowledgements

We thank Dr. Matteo Aldeghi, Dr. Robert Pollice, Dr. Mario Krenn and AkshatKumar Nigam for helpful discussions. R.J.H. gratefully acknowledges the Natural Sciences and Engineering Research Council of Canada (NSERC) for provision of the Postgraduate Scholarships-Doctoral Program (PGSD3-534584-2019). A. A.-G. acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. A. A.-G. also thanks Anders G. Frøseth for his generous support. We acknowledge the Defense Advanced Research Projects Agency (DARPA) under the Accelerated Molecular Discovery Program under Cooperative Agreement No. HR00111920027 dated August 1, 2019. The content of the information presented in this work does not necessarily reflect the position or the policy of the Government. All computations reported in this work are performed on the Niagara supercomputer at the SciNet HPC Consortium.[54,55] SciNet is funded by the Canada Foundation for Innovation, the Government of Ontario, Ontario Research Fund - Research Excellence, and by the University of Toronto.

References

(1) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679.

(2) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730.

(3) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 195–216.

(4) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3* (5), 5–20.

(5) Aspuru-Guzik, A.; Persson, K. *Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence.*; Mission Innovation: Innovation Challenge 6; Canadian Institute for Advanced Research, 2018.

(6) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Curr. Opin. Green Sustain. Chem.* **2020**, *25*, 100370.
(7) Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Synthesis of Many Different Types of Organic Small Molecules Using One Automated Process. *Science* **2015**, *347* (6227), 1221–1226.
(8) Mijalis, A. J.; Thomas, D. A.; Simon, M. D.; Adamo, A.; Beaumont, R.; Jensen, K. F.; Pentelute, B. L. A Fully Automated Flow-Based Approach for Accelerated Peptide Synthesis. *Nat. Chem. Biol.* **2017**, *13* (5), 464–466.
(9) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363* (6423).
(10) 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).
(11) Chatterjee, S.; Guidi, M.; Seeberger, P. H.; Gilmore, K. Automated Radial Synthesis of Organic Molecules. *Nature* **2020**, *579* (7799), 379–384.
(12) joseph, abragam; Pardo-Vargas, A.; Seeberger, P. H. Total Synthesis of Polysaccharides by Automated Glycan Assembly. *J. Am. Chem. Soc.* **2020**.
(13) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput. Aided Mol. Des.* **2007**, *21* (6), 311–325.
(14) Qiu, F. Strategic Efficiency — The New Thrust for Synthetic Organic Chemists. *Can. J. Chem.* **2008**, *86* (9), 903–906.
(15) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8.
(16) Fukunishi, Y.; Kurosawa, T.; Mikami, Y.; Nakamura, H. Prediction of Synthetic Accessibility Based on Commercially Available Compound Databases. *J. Chem. Inf. Model.* **2014**, *54* (12), 3259–3267.
(17) Berger, O.; Winters, K. R.; Sabourin, A.; Dzyuba, S. V.; Montchamp, J.-L. On the Cost of Academic Methodologies. *Org. Chem. Front.* **2019**, *6* (12), 2095–2108.
(18) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261.
(19) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**.
(20) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminformatics* **2020**, *12* (1), 35.
(21) Thakkar, A.; Chadimova, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) - Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. **2020**.
(22) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583.
(23) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443.
(24) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem. Int. Ed.* **2019**, *58* (14), 4515–4519.

(25) Chatterjie, N.; Stables, J. P.; Wang, H.; Alexander, G. J. Anti-Narcoleptic Agent Modafinil and Its Sulfone: A Novel Facile Synthesis and Potential Anti-Epileptic Activity. *Neurochem. Res.* **2004**, *29* (8), 1481–1486.
(26) Hou, W.; Bubliauskas, A.; Kitson, P.; Francoia, J.-P.; powell-Davies, H.; ParrillaGutierrez, J. M.; Frei, P.; Manzano, S.; Cronin, L. Automatic Generation of 3D Printed Reactionware for Chemical Synthesis Digitization Using ChemSCAD. **2020**.
(27) Jung, J.-C.; Lee, Y.; Son, J.-Y.; Lim, E.; Jung, M.; Oh, S. Simple Synthesis of Modafinil Derivatives and Their Anti-Inflammatory Activity. *Molecules* **2012**, *17* (9), 10446–10458.
(28) Fornaroli, M.; Velardi, F.; Colli, C.; Baima, R. Process for the Synthesis of Modafinil. US20040106829A1, June 3, 2004.
(29) Naddaka, V.; Menashe, N.; Lexner, J.; Saeed, S.; Kaspi, J.; Lerman, O. Process for the Preparation of Acetamide Derivatives. US20020183552A1, December 5, 2002.
(30) Maurya, S.; Yadav, D.; Pratap, K.; Kumar, A. Efficient Atom and Step Economic (EASE) Synthesis of the "Smart Drug" Modafinil. *Green Chem.* **2017**, *19* (3), 629–633.
(31) Lafon, L. Acetamide Derivatives. US4177290A, December 4, 1979.
(32) Prisinzano, T.; Podobinski, J.; Tidgewell, K.; Luo, M.; Swenson, D. Synthesis and Determination of the Absolute Configuration of the Enantiomers of Modafinil. *Tetrahedron Asymmetry* **2004**, *15* (6), 1053–1058.
(33) Kuehne, A. J. C.; Gather, M. C. Organic Lasers: Recent Developments on Materials, Device Geometries, and Fabrication Techniques. *Chem. Rev.* **2016**, *116* (21), 12823–12864.
(34) Gillis, E. P.; Burke, M. D. Multistep Synthesis of Complex Boronic Acids from Simple MIDA Boronates. *J. Am. Chem. Soc.* **2008**, *130* (43), 14084–14085.
(35) McCarthy, W. Z.; Corey, J. Y.; Corey, E. R. Group 4 9,10-Dihydro-9,10-Diheteroanthracenes: Synthesis and Structure. *Organometallics* **1984**, *3* (2), 255–263.
(36) Newman, S. G.; Aureggi, V.; Bryan, C. S.; Lautens, M. Intramolecular Cross-Coupling of Gem-Dibromoolefins: A Mild Approach to 2-Bromo Benzofused Heterocycles. *Chem. Commun.* **2009**, No. 35, 5236–5238.
(37) Wang, D.; Niu, Y.; Wang, Y.; Han, J.; Feng, S. Tetrahedral Silicon-Centered Imidazolyl Derivatives: Promising Candidates for OLEDs and Fluorescence Response of Ag (I) Ion. *J. Organomet. Chem.* **2010**, *695* (21), 2329–2337.
(38) Bunnett, J. F.; Zahler, R. E. Aromatic Nucleophilic Substitution Reactions. *Chem. Rev.* **1951**, *49* (2), 273–412.
(39) Ruiz-Castillo, P.; Buchwald, S. L. Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chem. Rev.* **2016**, *116* (19), 12564–12649.
(40) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365.
(41) Ou, Q.; Peng, Q.; Shuai, Z. Computational Screen-out Strategy for Electrically Pumped Organic Laser Materials. *Nat. Commun.* **2020**, *11* (1), 4485.
(42) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization for Categorical Variables Informed by Physical Intuition with Applications to Chemistry. *ArXiv200312127 Phys. Stat* **2020**.
(43) Fonin, A. V.; Sulatskaya, A. I.; Kuznetsova, I. M.; Turoverov, K. K. Fluorescence of Dyes in Solutions with High Absorbance. Inner Filter Effect Correction. *PLOS ONE* **2014**, *9* (7), e103878.
(44) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642–7655.
(45) Rohr, B.; Stein, H. S.; Guevarra, D.; Wang, Y.; Haber, J. A.; Aykol, M.; Suram, S. K.; Gregoire, J. M. Benchmarking the Acceleration of Materials Discovery by Sequential Learning. *Chem. Sci.* **2020**, *11* (10), 2696–2706.

(46) Felton, K.; Rittig, J.; Lapkin, A. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. **2020**.
(47) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. Olympus: A Benchmarking Framework for Noisy Optimization and Experiment Planning. *ArXiv201004153 Phys. Stat* **2020**.
(48) Gongora, A. E.; Xu, B.; Perry, W.; Okoye, C.; Riley, P.; Reyes, K. G.; Morgan, E. F.; Brown, K. A. A Bayesian Experimental Autonomous Researcher for Mechanical Design. *Sci. Adv.* **2020**, *6* (15), eaaz1708.
(49) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241.
(50) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3* (12), 1337–1344.
(51) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.* **2020**, *6* (20), eaaz8867.
(52) The Big Mac index https://www.economist.com/big-mac-index (accessed Jul 14, 2021).
(53) Seifrid, M.; Hickman, R. J.; Aguilar-Granda, A.; Lavigne, C.; Vestfrid, J.; Wu, T. C.; Gaudin, T.; Hopkins, E. J.; Aspuru-Guzik, A. Code and Data for "Routescore: Punching the Ticket to More Efficient Materials Development." Zenodo July 15, 2021.
(54) Loken, C.; Gruner, D.; Groer, L.; Peltier, R.; Bunn, N.; Craig, M.; Henriques, T.; Dempsey, J.; Yu, C.-H.; Chen, J.; Dursi, L. J.; Chong, J.; Northrup, S.; Pinto, J.; Knecht, N.; Zon, R. V. SciNet: Lessons Learned from Building a Power-Efficient Top-20 System and Data Centre. *J. Phys. Conf. Ser.* **2010**, *256*, 012026.
(55) Ponce, M.; van Zon, R.; Northrup, S.; Gruner, D.; Chen, J.; Ertinaz, F.; Fedoseev, A.; Groer, L.; Mao, F.; Mundim, B. C.; Nolta, M.; Pinto, J.; Saldarriaga, M.; Slavnic, V.; Spence, E.; Yu, C.-H.; Peltier, W. R. Deploying a Top-100 Supercomputer for Large Parallel Workloads: The Niagara Supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*; PEARC '19; Association for Computing Machinery: New York, NY, USA, 2019; pp 1–8.

Routescore: Punching the Ticket to More Efficient Materials Development

Martin Seifrid[a,b], Riley J. Hickman[a,b], Andrés Aguilar-Granda[a,b], Cyrille Lavigne[b], Jenya Vestfrid[a,b], Tony C. Wu[a,b], Théophile Gaudin[b,c], Emily J. Hopkins[a], Alán Aspuru-Guzik[a,b,d,e]

[a]*Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada*
[b]*Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada*
[c]*IBM Research Zürich, 8803 Rüschlikon, Zürich, Switzerland*
[d]*CIFAR Artificial Intelligence Research Chair, Vector Institute, Toronto, ON M5S 1M1, Canada*
[e]*Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, ON M5S 1M1, Canada*

Supplementary Information

Methods

RouteScore calculations

The *RouteScore* for each molecule was calculated with a set of custom Python scripts, which rely on an inventory file and a number of .pkl files that contain details of the hourly costs, reagents, t_H, t_M, etc.

All monetary quantities are given in Canadian dollars (CAD). Material costs were determined based on prices in CAD listed on supplier websites (usually Sigma Aldrich). In cases where prices were not available in CAD, the price was converted from US dollars (USD) by the following rough estimate: CAD = 1.33 × USD.

Estimates for the required human time for reactions were provided by a PhD-level chemist with extensive experience in synthetic organic chemistry. The t_H is based on estimates of how long the basic operations (setting up and monitoring the reaction, work-up, purification, cleaning up glassware, etc.) of each reaction would take the average chemist.

Estimating the labor cost of human and machine chemists

Table S1. Basis for estimates of hourly cost of performing iterative Suzuki-Miyaura cross-coupling with the Chemspeed automated synthesis platform.

General Parameters	**Quantity**	**Unit**
Parallel reactions	48	
Hours per reaction set	28	
Vials per reaction	5	
Biotage cartridges per reaction	2	
Solvent per reaction mixture and work up	0.02	L
Solvent of washing per reaction mixture	0.02	L
Cost breakdown	**CAD/unit**	**CAD/reaction**
Biotage cartridge	$1.76	$3.52
THF	$105.33	$4.21
Vials	$0.45	$2.26
Total costs	**CAD/hour**	**CAD/reaction**

	$17.13	$9.99

Table S2. Basis for estimates of hourly cost of human researchers.

General Parameters	**Hours**
Work weeks per year	50
Work hours per week	40
Cost breakdown	**CAD**
Salary (postdoctoral fellow)	$64,800.00
Benefits (10%)	$6,480.00
Overhead (53.5%)	$34,668.00
Total costs	**CAD**
Per year	$105,948.00
Per hour	$52.97

High-throughput quantum chemical calculations of molecular optical properties

We use a computational workflow consisting of semi-empirical, quantum chemical, and in-house software components to generate absorption and emission spectra and fluorescence rates for each molecule. In total, we subject 3458 molecules accessible through 3 general reaction types and 46 total fragments (Figure S1) to computation. We generate molecules from fragments using an in-house code based on the functionality of the RDKit.[1] Molecular conformers are generated using OpenBabel[2,3] and optimized using the xTB-GFN2 semi-empirical Hamiltonian.[4] The molecular Hessian is evaluated using the same approach at those geometries. Time-dependent density functional theory calculations (B3LYP/6-31G*) are then performed using Q-Chem 5.2[5] to obtain energy gradients for ground and excited electronic states. These gradients, projected over normal modes of the Hessian matrix, are used to estimate ground and excited state optimal geometries, using the Vertical Gradient method described in Ref. [6]. Normal modes of vibration and estimated optimal geometries were used to simulate absorption and emission spectra as well as fluorescence rates using standard path-integral equations.[7] We applied inhomogeneous broadening to the spectra (300 cm^{-1}).

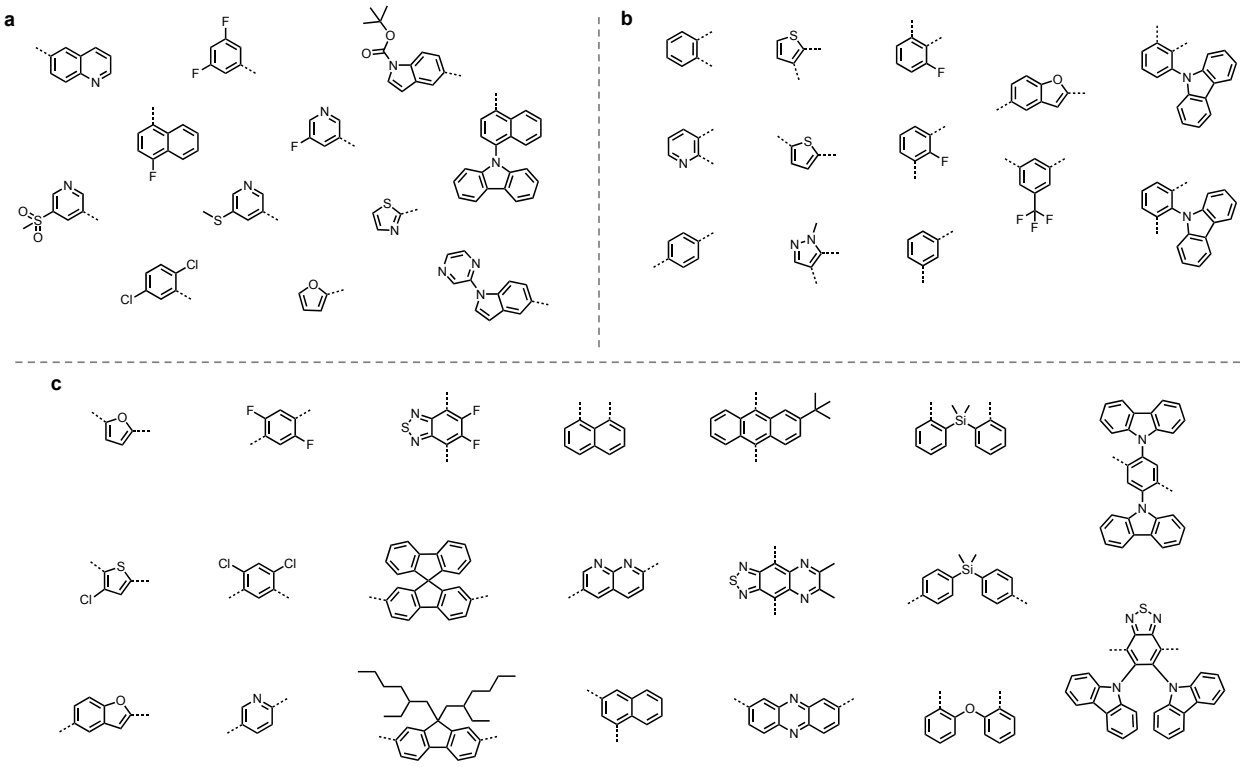

Figure S1. List of all fragments for computational multi-objective optimization.

Workflow results are post-processed to obtain the fluorescence peak score and the spectral overlap. The dimensionless fluorescence peak score is the fraction of the fluorescence power spectral density $\mathrm{PSD}(\omega)$ that falls within the interval from 400-460 nm, computed using the following formula,

$$\text{peak score} = \int_{U}^{L} \mathrm{d}\omega \mathrm{PSD}(\omega) \Big/ \int_{-\infty}^{\infty} \mathrm{d}\omega \mathrm{PSD}(\omega) \tag{S1}$$

where L and U are the frequencies of 460 nm and 400 nm light, respectively. The peak score is unity if all of the emitted light is within the desired wavelength interval, and is zero if all the emitted light falls outside of the desired interval.

The spectral overlap between emission and absorption is given by their L^2 inner product

$$\text{overlap} = \int_{-\infty}^{\infty} \mathrm{d}\omega \mathrm{PSD}(\omega)\varepsilon(\omega) \Big/ \sqrt{\int_{-\infty}^{\infty} \mathrm{d}\omega |\mathrm{PSD}(\omega)\varepsilon(\omega)|^2 \int_{-\infty}^{\infty} \mathrm{d}\omega |\varepsilon(\omega)|^2} \tag{S2}$$

where $\varepsilon(\omega)$ is the extinction coefficient. It is dimensionless and bound from below by zero and above by unity.

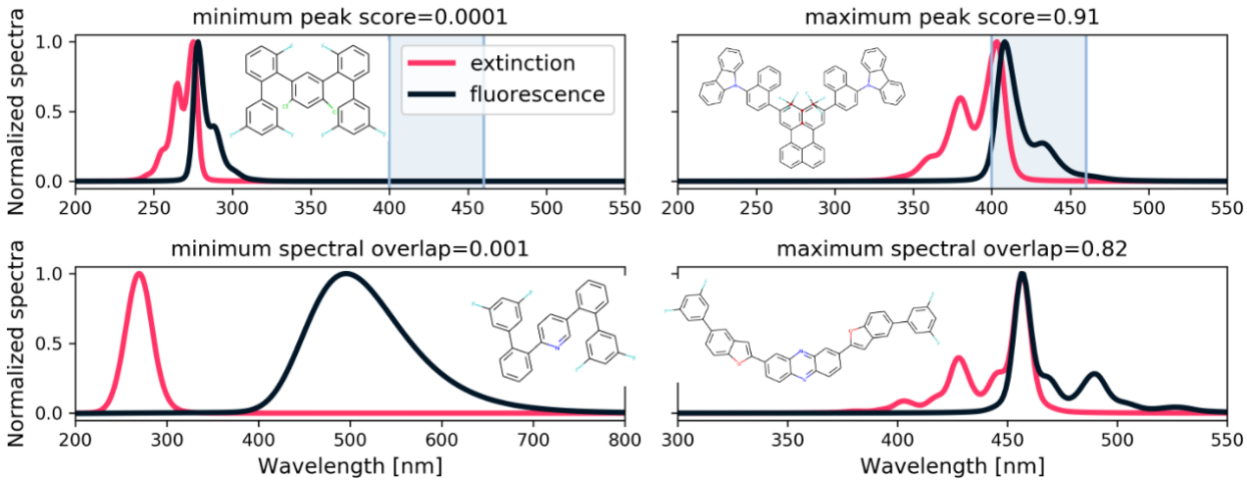

Figure S2. Visualization of the extreme values of the peak score and spectral overlap within the 3458 molecules considered in the inverse design optimization experiment. The region shaded in blue indicates the fluorescence spectrum target range (400-460 nm) used to compute peak score.

Multi-objective optimization of organic laser dye molecules

The optimization of the computational properties of laser dye molecules is framed as a multi-objective categorical optimization problem. There are three categorical variables, corresponding to the "A" , "B", and "C" fragments, with 14, 13 and 19 options respectively. We choose Bayesian optimization as the means of traversing the molecular space in search of desired target properties. Specifically, we employ the recently reported deep categorical Bayesian optimizer Gryffin.[8] Gryffin provides favorable scaling compared to Gaussian process-based strategies. We use the naive Gryffin implementation which does not consider physicochemical descriptors of the categorical options. Use of descriptors has been shown to accelerate the optimization rate of the algorithm and should be explored for this problem in future work. We use 2 sampling strategies, $\lambda = \{-1, 1\}$.

We consider 4 objectives: fluorescence peak score, *RouteScore*, spectral overlap, and fluorescence rate. *RouteScore* is calculated as described in the main text, while the other 3 are derived from the results of high-throughput quantum chemical calculations (See previous subsection). We use the scalarizing function Chimera[9] to transform sets of properties to a scalar-valued merit which is optimized using Gryffin. Chimera expects the objectives to be organized in a hierarchy representing the ranking of the objective's importances to the research goal. We also must provide tolerances to Chimera for each objective. Tolerances indicate the threshold value beyond which the researcher is satisfied with that objective's value, and moves on to optimize the subsequent objective in the hierarchy. We choose to use absolute tolerances in this study. The softness parameter of Chimera is set to 0.001.

Synthetic route optimization for a well-studied drug molecule

Table S3. References for each of the modafinil synthetic routes.

Route	Reference
1	Naddaka, *et al.*[10]
2	Maurya, *et al.*[11]
3	Maurya, *et al.*[11]
4	Chatterjie, *et al.*[12]
5	Hou, *et al.*[13]

6	Prisinzano, *et al.*[14]
7	Jung, *et al.*[15]
8	Fornaroli, *et al.*[16]
9	Lafon[17]
10	Lafon[17]

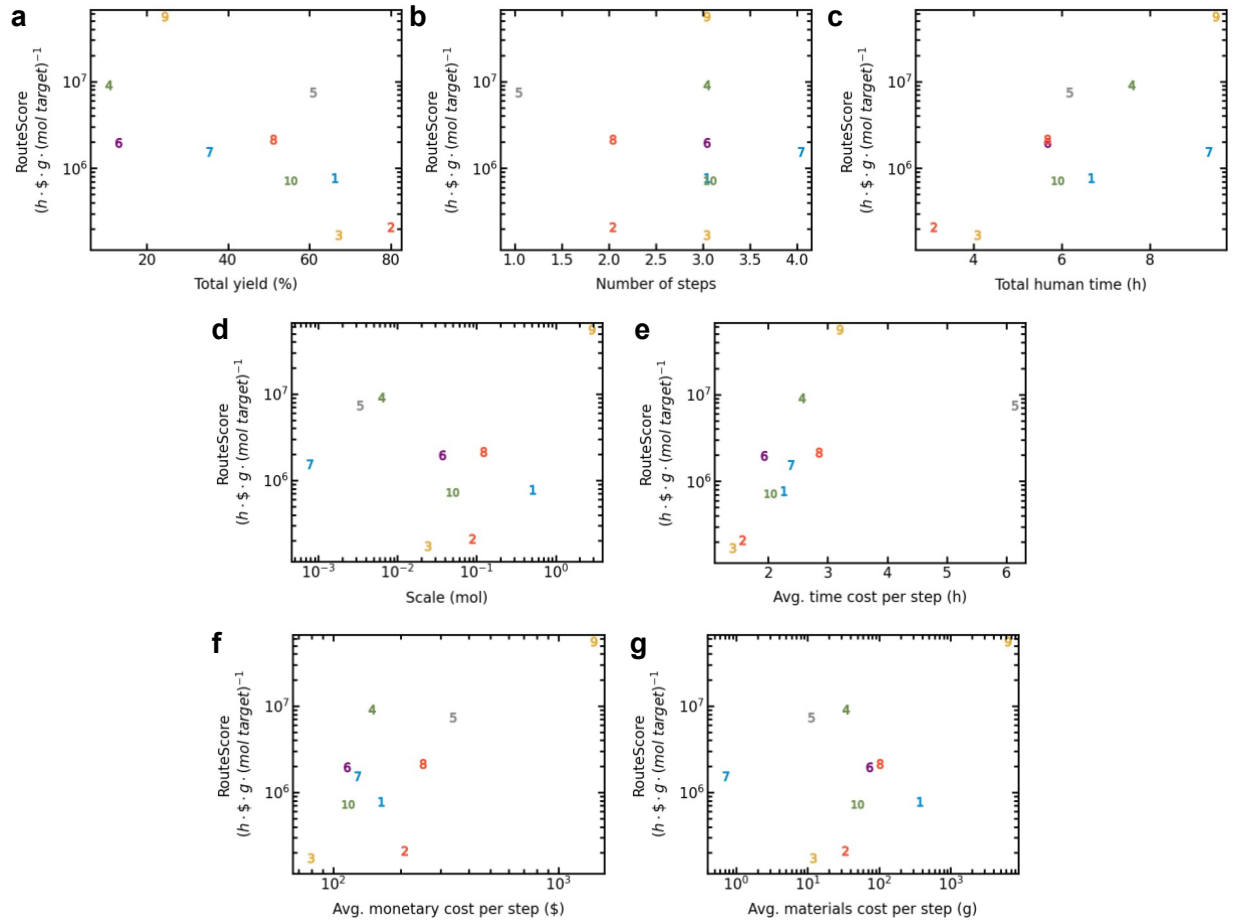

Figure S3. Correlation (or lack thereof) between *RouteScore* and (a) overall yield, (b) number of steps, (c) total human labor time, (d) scale, (e) average human time per step, (f) average monetary cost per step and (g) average mass cost per step for the 10 modafinil routes.

Table S4. Time estimates for the steps of each route to make modafinil.

Route	Step	Time (hours)	Time (h:mm)
1	1	1.5	1:30
	2	2.4167	2:25
	3	2.667	2:40
2	1	1	1:00
	2	2	2:00
3	1	1	1:00
	2	1.5	1:30
	3	1.5	1:30
4	1	2.75	2:45
	2	3.0833	3:05

	3	1.667	1:40
5	1	6.0833	6:05
6	1	1.667	1:40
	2	2.667	2:40
	3	1.25	1:15
7	1	3	3:00
	2	1.5	1:30
	3	3.5	3:30
	4	1.25	1:15
8	1	3.75	3:45
	2	1.833	1:50
9	1	4.833	4:50
	2	2.0833	2:05
	3	2.5	2:30
10	1	1.5	1:30
	2	2.4167	2:25
	3	1.833	1:50

Multi-objective optimization of organic laser molecules

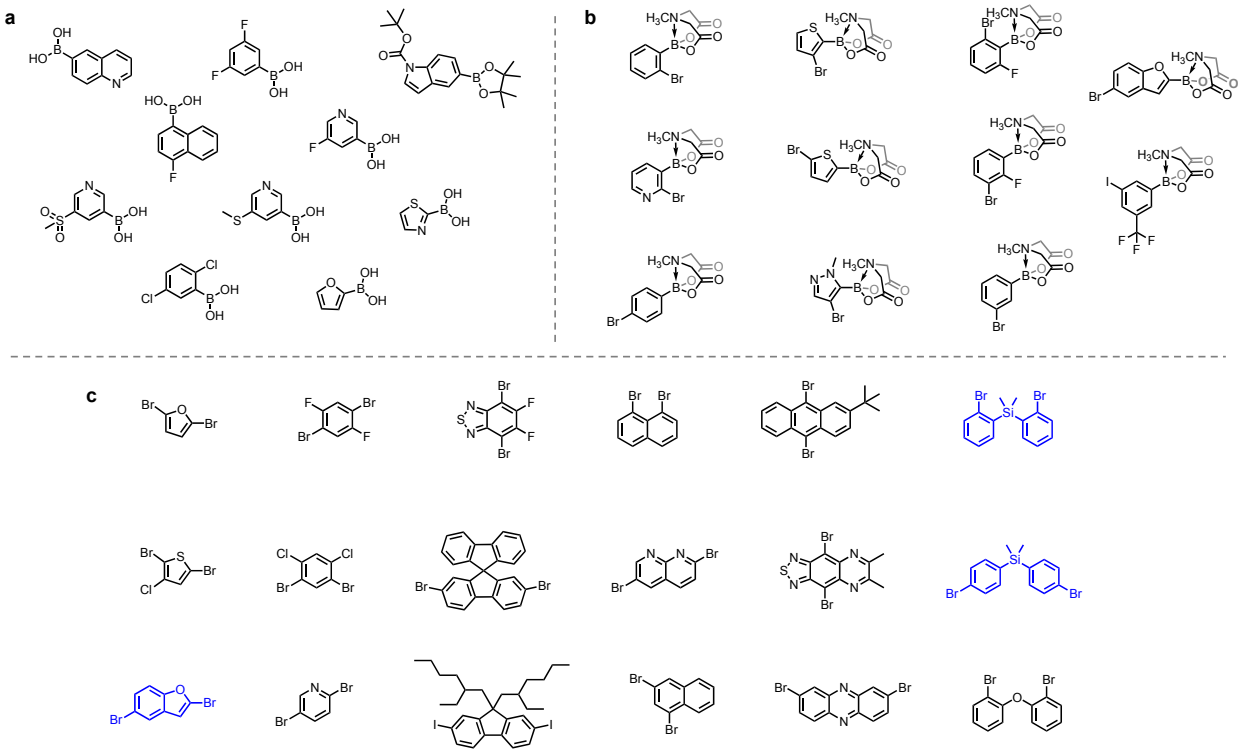

Figure S4. There are 10 "A" blocks (a), 11 "B" blocks (b) and 18 "C" blocks (c) that make up the space of 1980 symmetric pentamers that can be accessed by automated iterative Suzuki-Miyaura cross-coupling. The structures depicted in blue are not commercially available and must be manually synthesized.

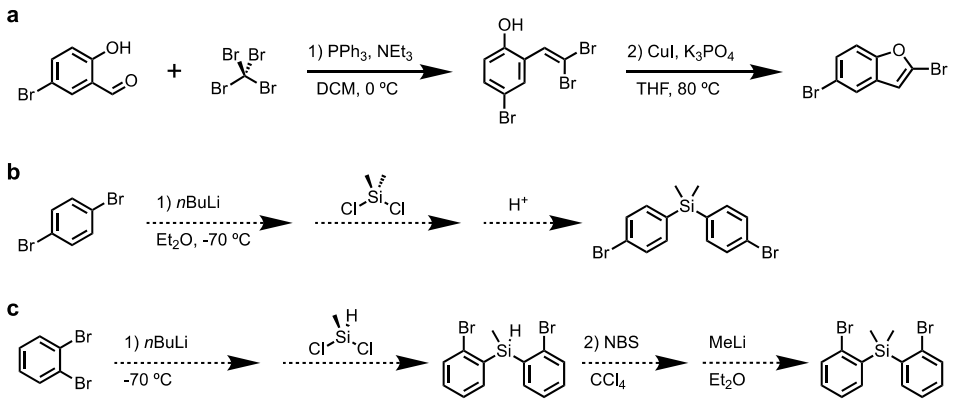

Figure S5. Synthetic routes for manually-synthesized building blocks: (a) 2,5-dibromobenzofuran, (b) bis(4-bromophenyl)dimethylsilane and (c) bis(2-bromophenyl)dimethylsilane. Dashed arrows represent one-pot multistep reactions, which are treated as a single step.

Table S5. Time estimates for manual syntheses. The steps correspond to those enumerated in Figure S5.

Reaction		Estimated labor (h)
2,5-Dibromobenzofuran	step 1	7.84
	step 2	2.52
Bis(4-bromophenyl)dimethylsilane	step 1	6.75
Bis(2-bromophenyl)dimethylsilane	step 1	6.05
	step 2	6.325
Boc deprotection		6.5
Buchwald-Hartwig amination		6.0
Nucleophilic aromatic substitution (S_NAr)		6.5

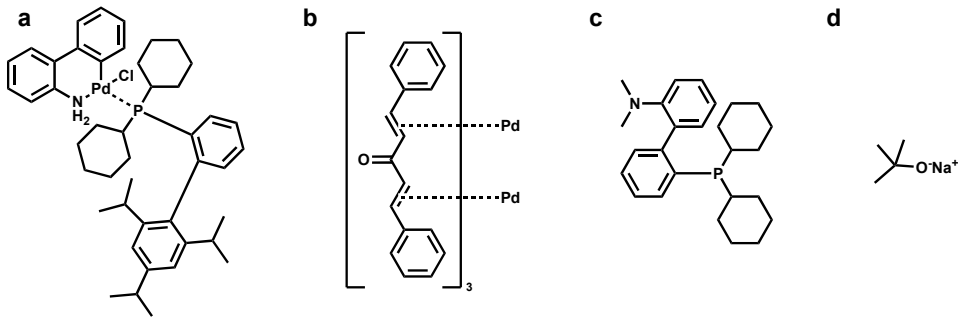

Figure S6. Molecular structures of the catalysts and ligands mentioned in the caption of Figure 5: (a) XPhos Pd G2 (chloro(2-dicyclohexylphosphino-2′,4′,6′-triisopropyl-1,1′-biphenyl)[2-(2′-amino-1,1′-biphenyl)]palladium(II)), (b) $Pd_2(dba)_3$ (tris(dibenzylideneacetone)dipalladium(0)), (c) DavePhos (2-dicyclohexylphosphino-2′-(*N*,*N*-dimethylamino)biphenyl), (d) NaO*t*-Bu (sodium *tert*-butoxide).

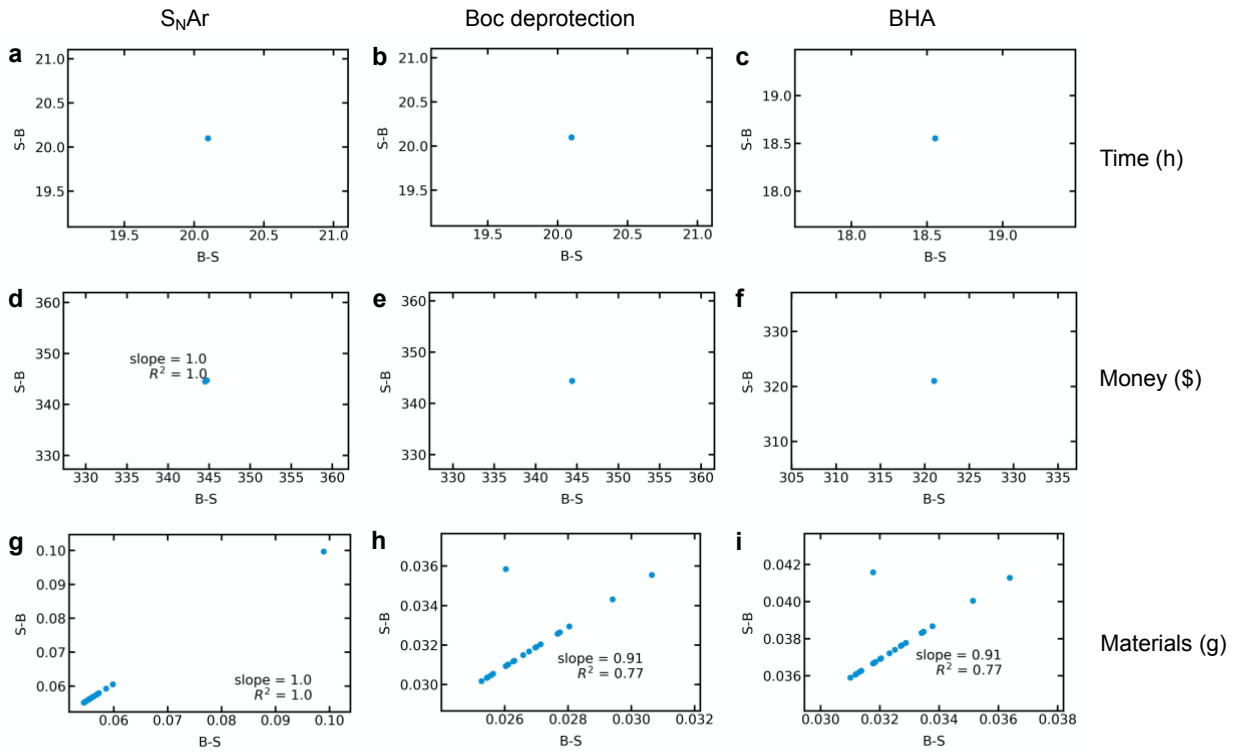

Figure S7. Analysis of the difference in costs of the S-B and B-S routes. Values for the S-B routes are in the vertical axis, while values for the B-S routes are in the horizontal axis. Rows correspond to the different cost components: time (a-c), money (d-f) and materials (g-i). Columns correspond to the different reactions: SNAr (a, d, g), Boc-deprotection (b, e, h) and BHA (c, f, i). The outliers in g, h and i are due to molecules with 4 carbazole groups instead of 2.

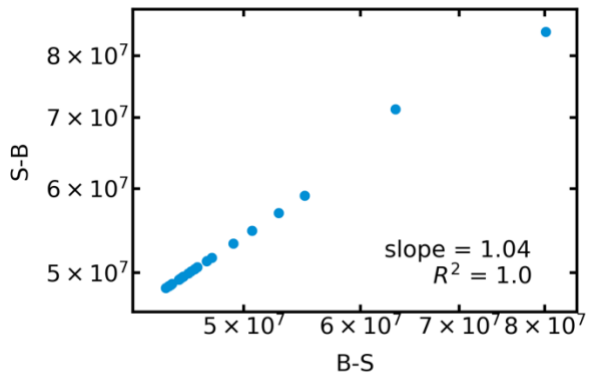

Figure S8. Comparison of *RouteScores* for the S-B and B-S routes. The S-B routes are systematically 4% more costly.

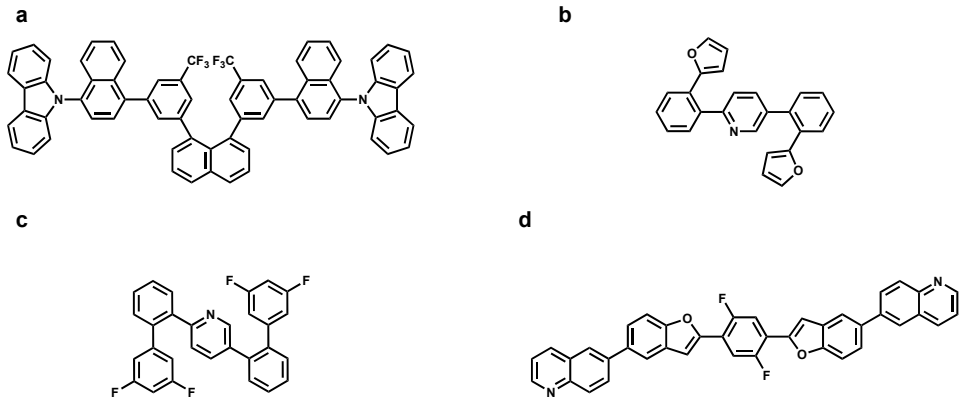

Figure S9. Structures of the molecules with (a) the best (highest) peak score, (b) the best (lowest) *RouteScore*, (c) the best (lowest) spectral overlap and (d) the best (highest) fluorescence rate.

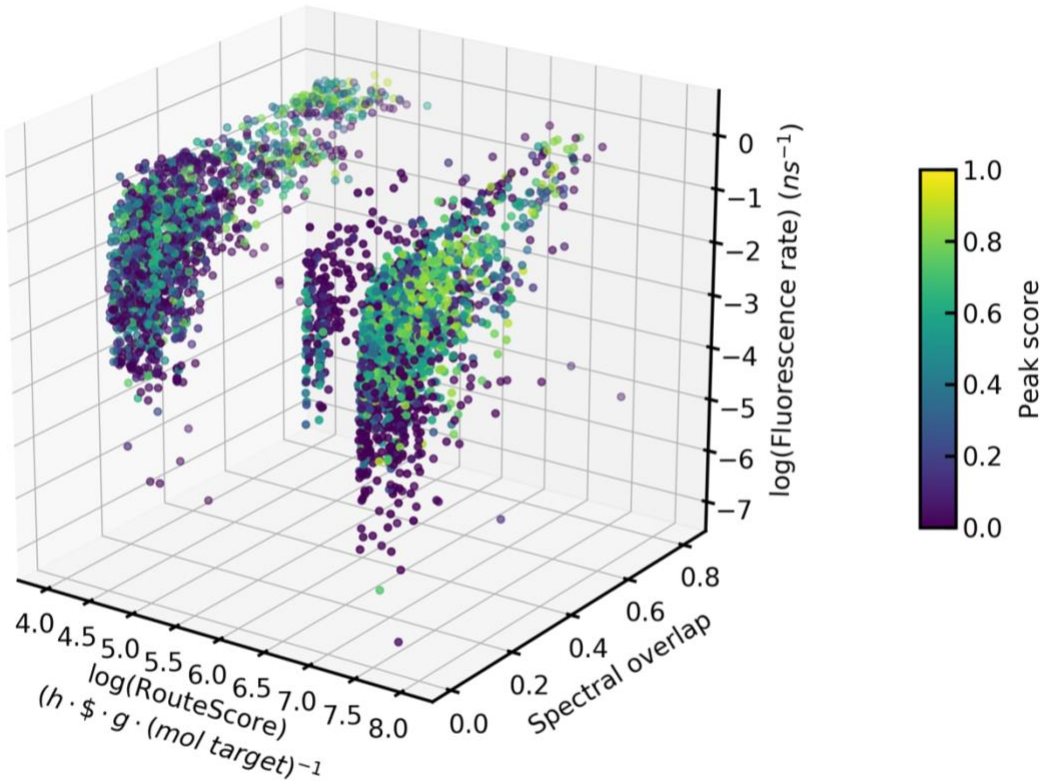

Figure S10. Distribution of the 4 properties of all 3458 molecules in the full molecular space.

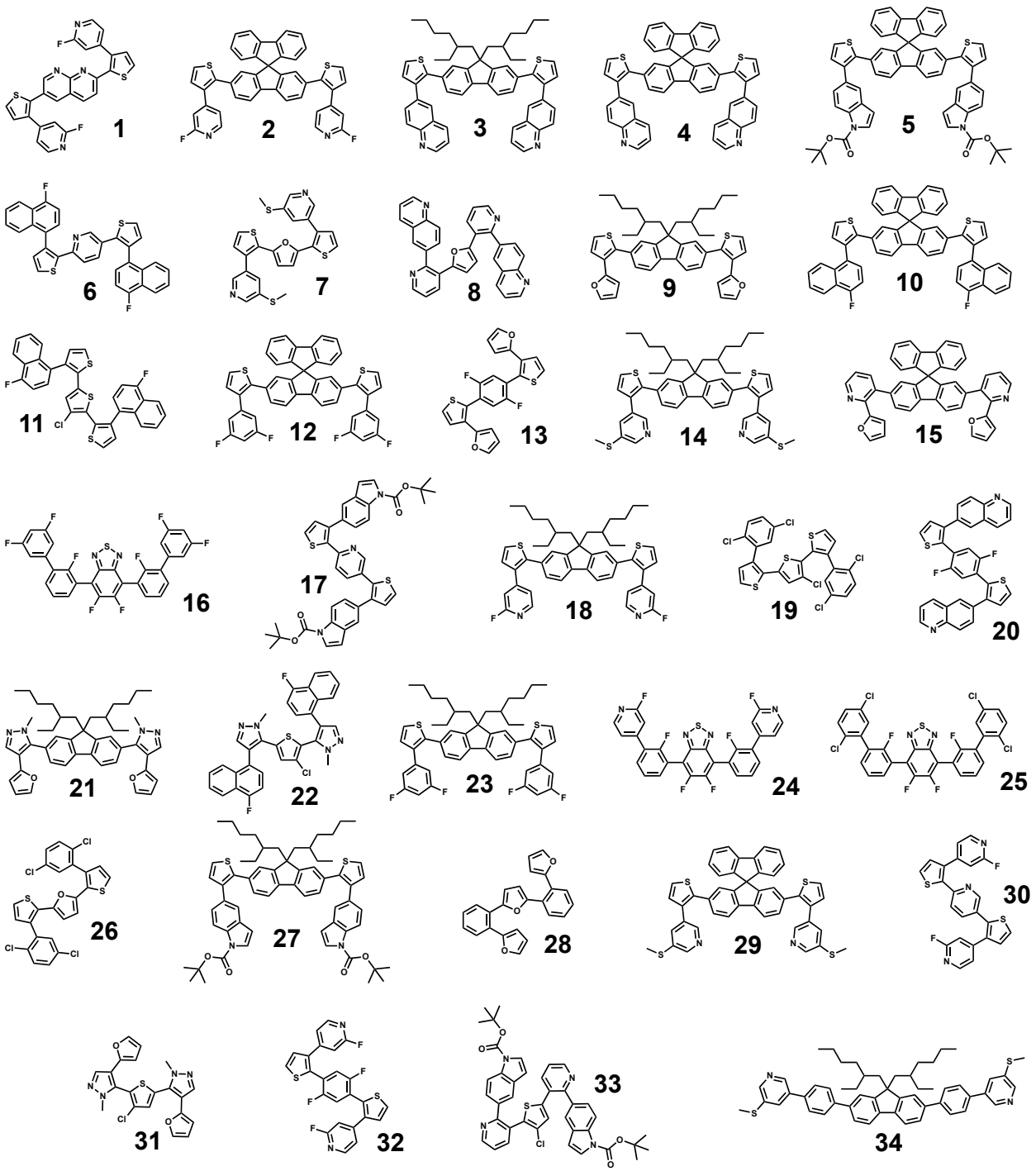

Figure S11. The 34 molecules that satisfied all 4 objectives, ordered from highest (1) to lowest (34) by merit.

Table S6. Properties of the 34 molecules in Figure S11.

Molecule rank	Peak score	*RouteScore*	Spectral overlap	Fluorescence rate (ns^{-1})
1	0.7934	60357.1	0.1115	0.4516
2	0.7827	58576.9	0.1039	0.4949

3	0.7772	58126.4	0.1526	0.6280
4	0.7763	58299.6	0.1908	0.4986
5	0.7674	87094.4	0.1670	0.6922
6	0.7641	68333.6	0.0904	0.2540
7	0.7638	97761.9	0.1319	0.3572
8	0.7616	57009.4	0.1411	0.3674
9	0.7568	39307.1	0.0935	0.5277
10	0.7501	69953.5	0.1450	0.5669
11	0.7493	71612.5	0.0851	0.2856
12	0.7480	41279.8	0.0480	0.5902
13	0.7459	38016.0	0.0445	0.1659
14	0.7457	98414.8	0.0972	0.6371
15	0.7447	39014.6	0.0493	0.2121
16	0.7395	41115.0	0.0788	0.1920
17	0.7331	85376.1	0.0424	0.3794
18	0.7326	58442.4	0.0340	0.5583
19	0.7325	46764.2	0.0450	0.3513
20	0.7315	56825.5	0.0253	0.1777
21	0.7307	39051.3	0.1554	0.2086
22	0.7304	71276.8	0.0767	0.1673
23	0.7286	41124.8	0.0697	0.6272
24	0.7283	58604.0	0.0679	0.1920
25	0.7255	44935.2	0.0793	0.1895
26	0.7242	44282.2	0.1186	0.4287
27	0.7187	86815.1	0.0776	0.5641
28	0.7150	37364.1	0.0334	0.2463
29	0.7139	98583.1	0.1070	0.5460
30	0.7100	57024.8	0.0237	0.4267
31	0.7063	40485.7	0.0912	0.1910
32	0.6992	57146.6	0.0165	0.3154
33	0.6894	88451.2	0.0144	0.1850
34	0.6881	99113.9	0.0309	1.0870

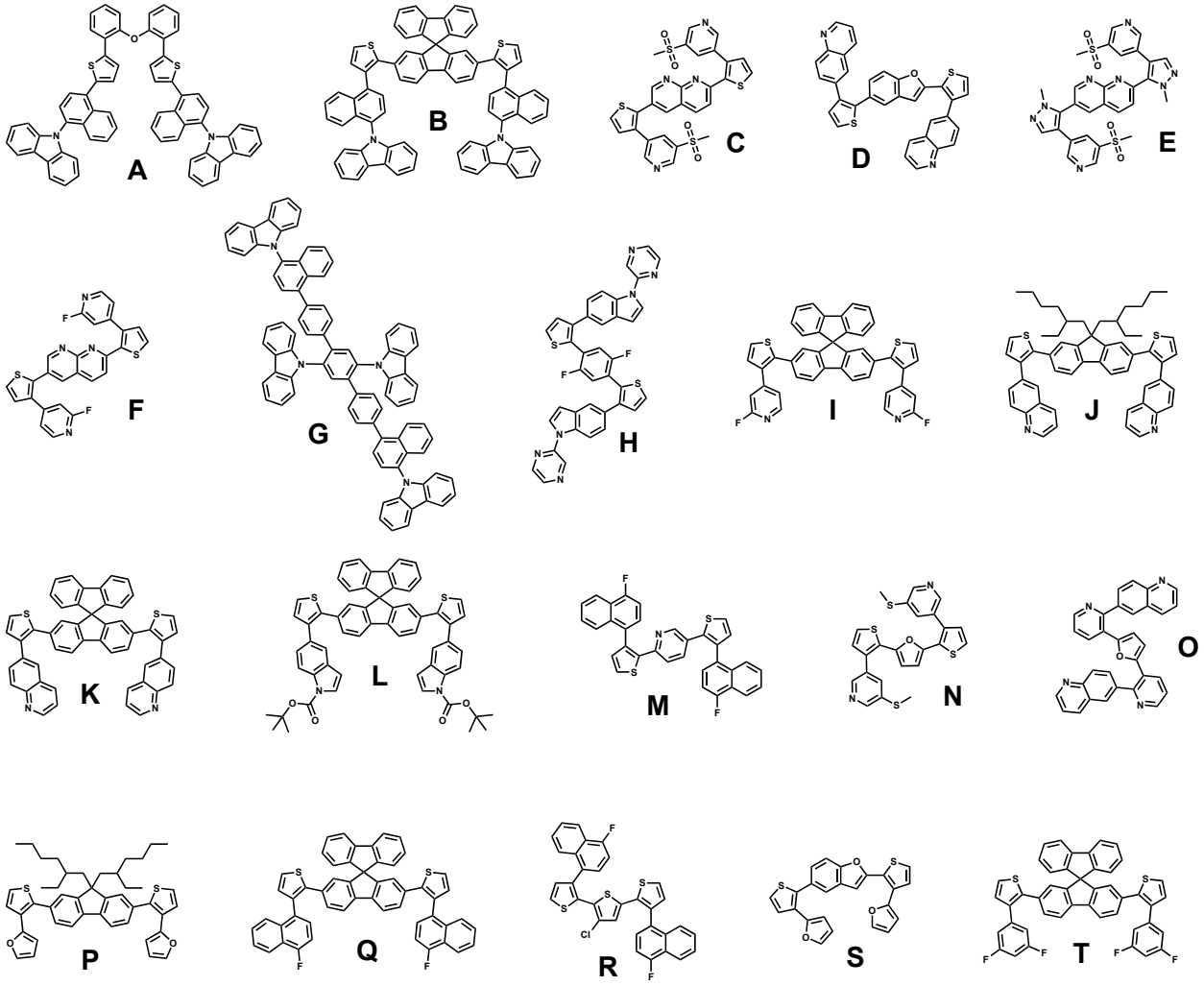

Figure S12. Top 20 molecules by merit in the three-objective hierarchy (i.e., without optimizing the *RouteScore*), ordered from best (A) to worst (T) by merit.

Table S7. Properties of the molecules in Figure S12.

Molecule rank	Peak score	*RouteScore*	Spectral overlap	Fluorescence rate (ns^{-1})
A	0.8598	22550876.1	0.1937	0.2071
B	0.8298	23563142.3	0.1287	0.3195
C	0.8291	208347.1	0.1709	0.4555
D	0.8274	34944339.2	0.1604	0.2397
E	0.8208	207807.4	0.1864	0.1586
F	0.7934	60357.1	0.1115	0.4516
G	0.7926	40271200.7	0.1473	0.2537
H	0.7863	21943307.3	0.0981	0.2834
I	0.7827	58576.9	0.1039	0.4949
J	0.7772	58126.4	0.1526	0.6280
K	0.7763	58299.6	0.1908	0.4986

L	0.7674	87094.4	0.1670	0.6922
M	0.7641	68333.6	0.0904	0.2540
N	0.7638	97761.9	0.1319	0.3572
O	0.7616	57009.4	0.1411	0.3674
P	0.7568	39307.1	0.0935	0.5277
Q	0.7501	69953.5	0.1450	0.5669
R	0.7493	71612.5	0.0851	0.2856
S	0.7485	34925534.8	0.0627	0.2312
T	0.7480	41279.8	0.0480	0.5902

We observe systematic identification of molecules with larger fluorescence rates in the blue traces after roughly 150 evaluations compared to the maroon traces (Figure 7e). This could be caused by the fact there exists a slight negative monotonic correlation between the *RouteScore* and the fluorescence rates for the candidate molecules considered here (Pearson coefficient of –0.25, Figure S13). A possible explanation for this observation is that larger molecules fluoresce more strongly, but are typically more expensive to synthesize.

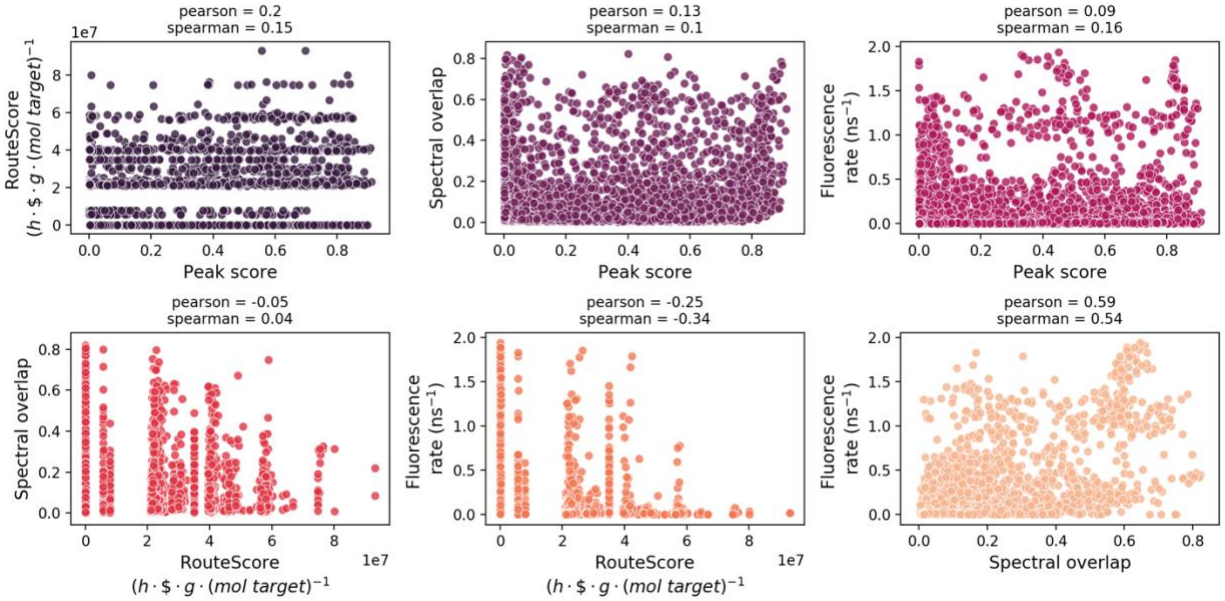

Figure S13. Pairwise correlations for all unique pairs of objectives calculated using the entire set of 3458 laser molecules. The fluorescence rate has a moderate positive linear correlation with the spectral overlap as evidenced by the Pearson correlation coefficient of 0.59. We also note a slight negative correlation between the fluorescence rate and the *RouteScore* (Pearson correlation coefficient of -0.25) for this set of molecules.

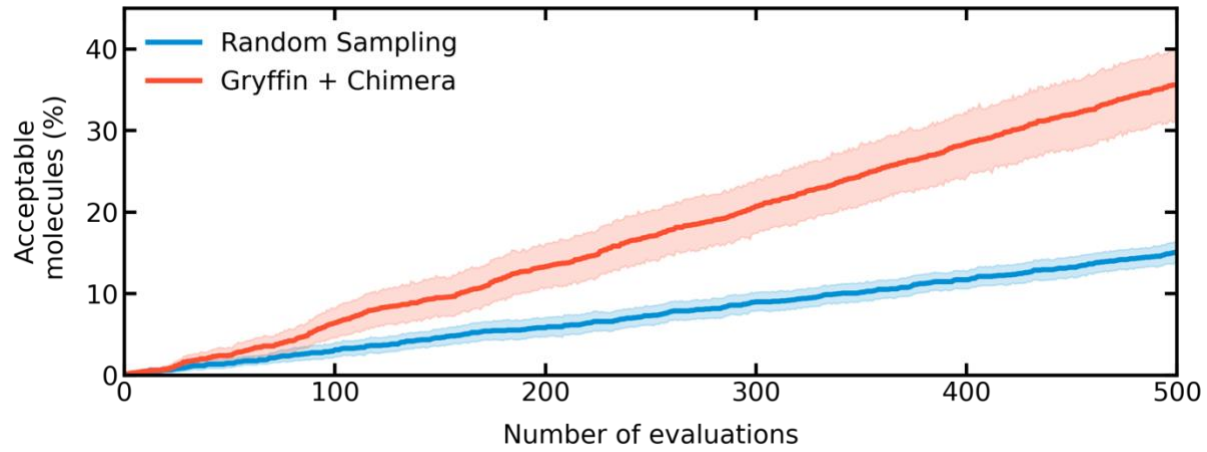

Figure S14. Percentage of satisfactory molecules identified by each strategy over 500 evaluations. Each trace comprises 40 independently seeded runs, and the shaded area displays a 95% confidence interval. The Gryffin + Chimera strategy identifies on average 35±3% of satisfactory molecules after 500 evaluations, while random sampling identifies 15±1% of satisfactory candidates.

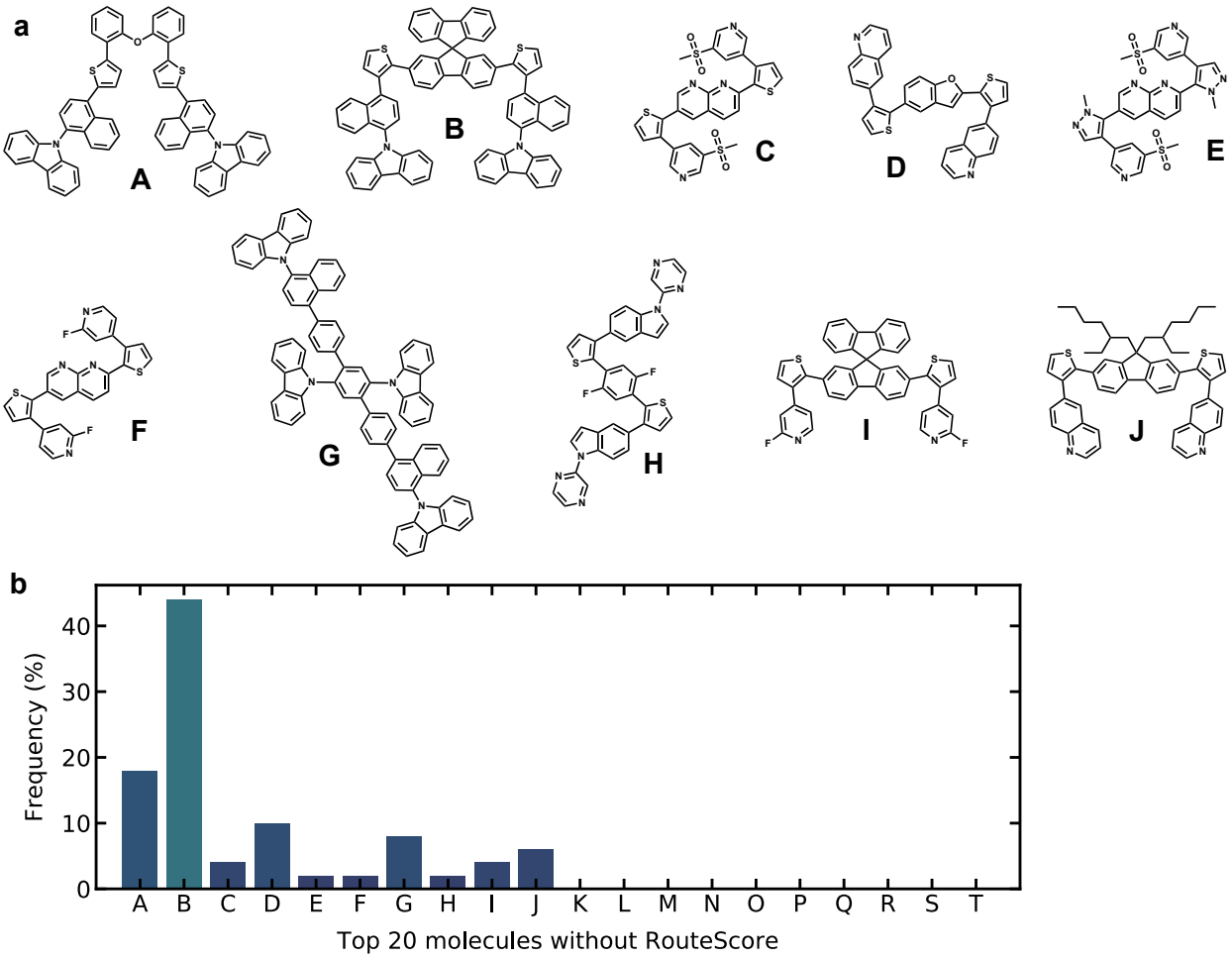

Figure S15. Structures (a) of the top three-objective molecules and the frequency with which the optimizations found them. Molecular structures of all top 20 three-objective molecules are provided in Figure S12.

References

(1) *RDKit: Open-Source Cheminformatics*; 2006.

(2) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33.

(3) *The Open Babel Package*.

(4) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671.

(5) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; III, H. L. W.; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; Jr, R. A. D.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; III, W. A. G.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; III, H. F. S.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Voorhis, T. V.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113* (2), 184–215.

(6) de Souza, B.; Neese, F.; Izsák, R. On the Theoretical Prediction of Fluorescence Rates from First Principles Using the Path Integral Approach. *J. Chem. Phys.* **2018**, *148* (3), 034104.

(7) Baiardi, A.; Bloino, J.; Barone, V. General Time Dependent Approach to Vibronic Spectroscopy Including Franck–Condon, Herzberg–Teller, and Duschinsky Effects. *J. Chem. Theory Comput.* **2013**, *9* (9), 4097–4115.

(8) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization for Categorical Variables Informed by Physical Intuition with Applications to Chemistry. *ArXiv200312127 Phys. Stat* **2020**.

(9) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642–7655.

(10) Naddaka, V.; Menashe, N.; Lexner, J.; Saeed, S.; Kaspi, J.; Lerman, O. Process for the Preparation of Acetamide Derivatives. US20020183552A1, December 5, 2002.
(11) Maurya, S.; Yadav, D.; Pratap, K.; Kumar, A. Efficient Atom and Step Economic (EASE) Synthesis of the "Smart Drug" Modafinil. *Green Chem.* **2017**, *19* (3), 629–633.
(12) Chatterjie, N.; Stables, J. P.; Wang, H.; Alexander, G. J. Anti-Narcoleptic Agent Modafinil and Its Sulfone: A Novel Facile Synthesis and Potential Anti-Epileptic Activity. *Neurochem. Res.* **2004**, *29* (8), 1481–1486.
(13) Hou, W.; Bubliauskas, A.; Kitson, P. J.; Francoia, J.-P.; Powell-Davies, H.; Gutierrez, J. M. P.; Frei, P.; Manzano, J. S.; Cronin, L. Automatic Generation of 3D-Printed Reactionware for Chemical Synthesis Digitization Using ChemSCAD. *ACS Cent. Sci.* **2021**.
(14) Prisinzano, T.; Podobinski, J.; Tidgewell, K.; Luo, M.; Swenson, D. Synthesis and Determination of the Absolute Configuration of the Enantiomers of Modafinil. *Tetrahedron Asymmetry* **2004**, *15* (6), 1053–1058.
(15) Jung, J.-C.; Lee, Y.; Son, J.-Y.; Lim, E.; Jung, M.; Oh, S. Simple Synthesis of Modafinil Derivatives and Their Anti-Inflammatory Activity. *Molecules* **2012**, *17* (9), 10446–10458.
(16) Fornaroli, M.; Velardi, F.; Colli, C.; Baima, R. Process for the Synthesis of Modafinil. US20040106829A1, June 3, 2004.
(17) Lafon, L. Acetamide Derivatives. US4177290A, December 4, 1979.