Computer#aided evaluation and exploration of chemical spaces constrained by reaction pathways

Citation: Levin I, Fortunato ME, Tan KL, Coley CW. Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways. AIChE J. 2023; 69(12):e18234.

As Published: 10.1002/aic.18234

Publisher: Wiley

Persistent URL: https://hdl.handle.net/1721.1/158182

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

Received: 14 June 2023 | Revised: 11 August 2023 | Accepted: 14 August 2023
DOI: 10.1002/aic.18234

AIChE JOURNAL

RESEARCH ARTICLE
Process Systems Engineering

Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways

Itai Levin[1] | Michael E. Fortunato[2] | Kian L. Tan[2] | Connor W. Coley[3,4]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Novartis Institutes for BioMedical Research, Cambridge, Massachusetts, USA
[3]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence
Connor W. Coley, Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.
Email: ccoley@mit.edu

Funding information
Machine Learning for Pharmaceutical Discovery and Synthesis consortium; National Institute of Allergy and Infectious Diseases, Grant/Award Number: R21AI169342

Abstract

The processes of molecular design and synthetic route selection are necessarily intertwined during discovery. Computational tools have been developed to facilitate synthesis planning, but in a discovery setting, finding a single route to a single molecule of interest may be less important than finding a route that enables rapid access to a library of analogs. Here, we demonstrate how we can estimate route "diversifiability" and use it as a criterion during route selection. We illustrate how the chemical space of synthetically accessible analogs is influenced by properties of alternative starting materials or constraints on their cost. Finally, we integrate these analyses with a synthesizability-constrained hit expansion workflow in a virtual screening pipeline for focused library expansion around putative hits to support molecular optimization. As medicinal chemistry and adjacent fields shift toward more autonomous design and synthesis of new molecules, it will be increasingly important to embed considerations of synthesizability into molecular design to ensure that computational recommendations are actionable.

KEYWORDS
cheminformatics, drug discovery, molecular design, synthesis planning

1 | INTRODUCTION

The synthesis of new molecules is an essential task during discovery and is often a practical bottleneck in the design-make-test-analyze cycle. In principle, at each iteration, we might have multiple molecules for which we must devise multiple distinct experimental synthesis strategies for testing. The ideation of synthetic routes for new molecules can be supported by computer-aided synthesis planning (CASP) tools, which have matured in recent years.[1] Nevertheless, the complexity of modern active pharmaceutical ingredients (APIs) and analogously complex preclinical candidates sometimes exceeds the capability of data-driven CASP tools; even if a viable route is identified, screening conditions and validating each step is burdensome.

"Make-on-demand" virtual chemical libraries mitigate the need to plan new synthetic routes. These libraries are typically defined by a set of versatile chemical reactions and a set of available chemical building blocks such that billions of molecules can be enumerated virtually, all of which we expect to be readily synthesizable.[2-5] Such libraries have existed within the firewalls of pharmaceutical companies for years, but have increased in accessibility recently due to commercial vendors like Enamine and WuXi.[6] They have become invaluable tools for drug discovery because any candidate hits identified *in silico* can be synthesized and tested *in vitro* with only a short lag time. Success stories of molecular discovery via these synthetically constrained virtual spaces in structure-based drug design include the discovery of new anti-depressants,[7] anti-inflammatories,[8] analgesics,[9] and antivirals.[10]

Virtual libraries built in a property-agnostic way are most useful for hit finding when combined with computational evaluation workflows (e.g., structure-based drug design techniques). At later stages of discovery, during iterative molecular optimization, our evaluations typically make use of physical experiments. We might have an idea of what molecule to make in the current design cycle, but we expect that in the next design cycle we will propose modifications to it in pursuit of an even-better candidate.[11] These modifications are often "local"—only minor modifications—for at least three reasons: (a) testing similar compounds from a single chemical family can reveal specific clues about favorable/unfavorable modifications à la matched molecular pair analysis;[12] (b) if we are using a surrogate property prediction model to guide generation, its domain of applicability might make it only reliable in the chemical space closely surrounding known (training) compounds; and (c) developing new syntheses can be expensive/difficult, so chemists have a natural preference towards what is feasible to make given a modest number of familiar synthetic strategies and minor changes thereto.

As a complement to make-on-demand approaches, more focused enumeration strategies can generate libraries of chemical analogs. This can be done with a fragment-based approach, wherein a molecule is decomposed into constituent substructures that are exchanged with other predefined substructures and recombined to form new molecules.[13-17] While this generates structurally related compounds, the synthesizability of the enumerated molecules is not guaranteed. Alternatively, some iterative approaches start with building blocks and a set of well-characterized chemical reactions (similar to make-on-demand libraries) and grow a reaction tree to explore the chemical space that can be reached from the building blocks with the reactions.[18-23] Every enumerated molecule is associated with a unique synthetic route, but the synthetic routes are not necessarily related to one another. Finally, tools such as PathFinder[24,25] and Synthesia[26] employ a reaction-based approach that explicitly constrains the synthesizability of the generated molecules. The enumeration takes as input a seed molecule, synthesis plan, and set of building blocks, and enumerates analogues by running different combinations of building blocks through the original synthesis plan such that enumerated molecules can theoretically all be synthesized using the same sequence of chemical reactions.

Using a route-based enumeration strategy to generate chemical libraries raises the question of how to prioritize candidate synthesis plans and arrive at the one used for enumeration. While elements of synthesis planning are frequently applied to the domain of molecular discovery, the reverse–using elements of the molecular discovery workflow to select synthesis plans–has not been explored. With the goal of maximizing the number of chemical analogs accessible with a single route in mind, a criterion for pathway selection is the extent to which a pathway is conducive to diversification. Diversification is an important consideration because the hit expansion and lead optimization processes, as alluded to previously, require the synthesis of many candidates en route to a final preclinical candidate. Understanding up front which routes lead to a larger accessible space can influence synthetic resource allocation and early selection of robust synthesis plans that facilitate access to a diverse chemical space can accelerate the discovery process.

In this article, we report an open-source computational workflow to (1) score and select retrosynthetic pathways on the basis of perceived route diversifiability, (2) estimate property distributions of the resulting enumerated space to inform or constrain reactant selection, and (3) enumerate synthetic pathway-constrained analogs via selection of suitable alternative reactants. We envision this tool will enable medicinal chemists to generate large virtual spaces ($> 10^6$), which is currently a skill set limited to computational chemists and cheminformaticians. The key premise of this article is that considering the ease of accessing many analogs during route selection is a more efficient strategy than decoupling molecular design and synthesis, which may create the need to revise the synthetic route entirely. We demonstrate this process on a hypothetical virtual screening pipeline and illustrate how we can improve upon properties of an initial hit compound through pathway-constrained hit expansion. This combination of our enumeration methodology with property models enables a readily accessible workflow for the design of new synthesizable molecules accessible to those without traditional computational expertise.

2 | SELECTING PATHWAYS USING A DIVERSIFIABILITY METRIC

2.1 | Estimating route diversifiability

Evaluating computationally generated synthesis plans in terms of quantitative criteria remains a challenge.[27] In the context of molecular discovery, the number and properties of analogous molecules that can be synthesized with one synthesis plan is an important consideration. Previous approaches to characterizing the accessible analog space for a given synthetic route rely on the explicit enumeration of the molecules.[24,26] However, explicit enumeration can be a computationally expensive procedure. Enumerated product spaces are combinatorially large with respect to the number of building blocks matched to the reaction plan. The cost of generating and storing enumerated spaces rapidly becomes nonnegligible once their size starts to exceed 10^6 distinct compounds. For this reason, we would like to avoid explicit enumeration of such spaces and instead perform as much of the analysis as possible with an *implicit* enumerated space when considering the set of building blocks that are deemed compatible with the reaction sequence.

We developed a workflow to estimate the number of product analogs that can be enumerated with a synthetic route and a set of predefined building blocks with no explicit enumeration (Figure 1A). We refer to the number of accessible analogs as the route's "diversifiability." To avoid enumeration, diversifiability is calculated as the number of combinations of molecules from the set of building blocks that can potentially be run in the forward direction through the synthetic route.

For an input route, a structural query pattern is defined for each starting material to capture which chemical moieties the molecule must contain to remain compatible with the synthetic route. This pattern is used to query the building block database for analogs to the

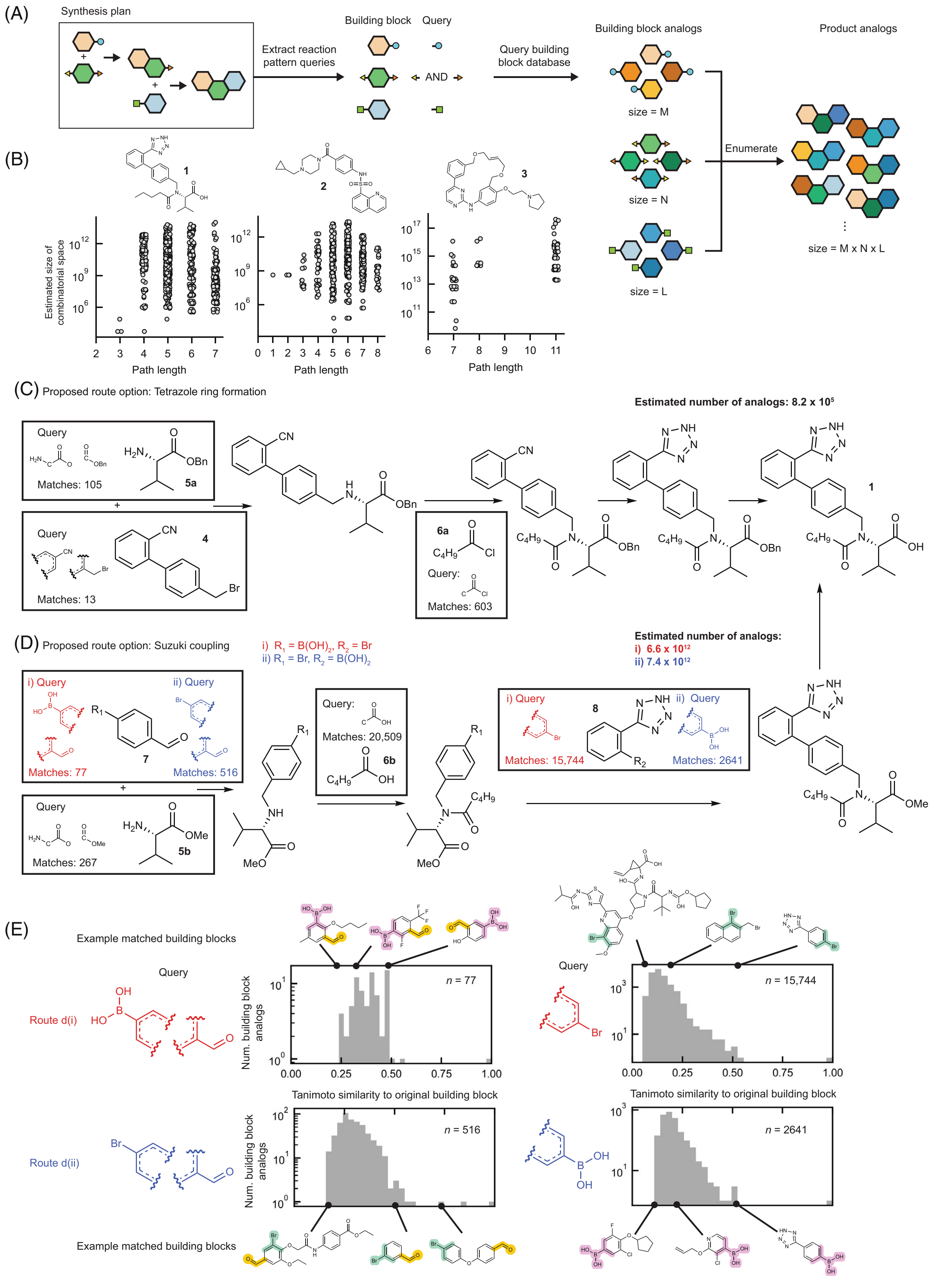

FIGURE 1 Legend on next page.

original starting materials. Examples of building blocks and their corresponding queries are shown for the routes in Figure 1C,D. The number of analogs for each starting material is counted. The route's diversifiability is the product of these counts.

We recognize that simple substructure matching is not a sufficient criterion for determining experimental substrate compatibility, but it serves as a useful first approximation in this context. Our approach is consistent with the use of algorithmically extracted reaction templates in retrosynthetic planning, which provide approximations of compatibility but still benefit from additional feasibility filters.[28,29]

This workflow is designed to be useful for users considering either a small set of manually curated routes or a large set of computer-generated routes. We expect most users to follow the former human-in-the-loop workflow at first and transition to a more automated CASP-driven workflow over time. The input is a synthetic route represented as a list of single-step reaction SMILES. These require little expertise to define using an interactive molecule editor such as ChemDraw or they can be automatically retrieved from the outputs of CASP software. The network structure of the synthetic route is automatically inferred. This obviates the need to input the route as a tree graph as in Synthesia, which can be cumbersome to define manually. Additionally, each reaction from the route is atom-atom mapped using RXNMapper[30] and generalized reaction SMARTS patterns are algorithmically defined using RDChiral.[31] This means that the platform is adaptable to new chemical reactions without requiring users to manually define any SMARTS patterns, effectively expanding the userbase of this methodology beyond SMARTS-fluent cheminformaticians.

2.2 | Diversifiability-informed route selection

The reduced computational cost of our route diversifiability calculation compared to an explicit enumeration of analog space makes it possible to estimate the size of the accessible space for a route in 1–100 CPU seconds with a set of approximately 10^6 building blocks, depending on the complexity of the route. This makes the analysis of thousands of proposed reaction sequences practical, rendering it feasible to use diversifiability as an additional metric for comparing synthetic routes proposed by a CASP tool on top of other commonly used metrics such as step count or longest linear sequence.

We queried the ASKCOS (v.2023.01)[29] tree-builder tool for synthesis routes to three FDA-approved drugs, valsartan (**1**), mitapivat (**2**), and pacritinib (**3**), and evaluated the diversifiability and pathway length for each returned route (Figure 1B). The search was limited to a maximum depth of 5 and a maximum search time of 180 s. The buyable compound dataset used for the search was left as its default of 280,469 compounds curated from the Sigma Aldrich, eMolecules, and WuXi Lab Network catalogs with list prices $<\$100/g$. Each generated route is a set of single-step reaction SMILES that lead from the buyable building blocks to the final product.

The relationship between path length and diversifiability is not monotonic. In some cases, a longer path corresponds to a larger number of starting materials for which there are more analogs in the building block database, leading to a greater number of product analogs. In other cases, the longer path introduces more constrained queries for building block compatibility, reducing the number of matched analogs. We find that the route diversifiability can vary by many orders of magnitude for a given target molecule, even for routes of the same length.

We highlight how estimates of diversifiability can be used to compare proposed synthesis routes using the angiotensin II receptor blocker, valsartan (**1**), as an illustrative example. First, we compare two plausible strategies proposed by ASKCOS: one where the biaryl core is introduced by a starting material (**4**), and the tetrazole ring is formed via condensation of a nitrile group and an azide salt[32] in the final step (Figure 1C) and one where the the biaryl structure is formed via a Suzuki coupling in the last step of the synthesis (Figure 1D). The estimated diversifiability of the Suzuki route is over 10^6 times greater than the diversifiability of the route that proceeds via tetrazole ring formation.

The discrepancy in the size of the accessible chemical space between the two strategies can be better understood by looking at the number of analog matches for each starting material. Few building blocks in the database contain both an aryl nitrile and a benzylic bromide compared to the great number of building blocks that contain the aryl boronic acid or aryl bromide required for the Suzuki reaction. There are only 13 matches for the former compared to over 1,000,000 combinations of the latter. Other more minor differences account for the rest of the discrepancy. The methyl protected amino acid (**5b**) has approximately two times more matches than the benzyl protected amino acid (**5a**) and the acid (**6b**) has approximately 34 times more matches than the acyl chloride (**6a**).

The Suzuki coupling can hypothetically be performed with the aryl boronic acid and the aryl halide on either reactant molecule. The ASKCOS tree builder returns both options. We sought to understand how swapping which reactant had which moiety could affect the diversifiability of the route. The route with the bromide on **7** and the boronic acid on **8** (Figure 1D(ii)) is estimated to lead to a chemical space that contains approximately 8×10^{11} more molecules (12% increase) than the route with the opposite configuration (Figure 1D(i)) Were our goal the selection of a route that enables the greatest

FIGURE 1 (A) Schematic of reaction-based library enumeration, where required reactive patterns for each starting material are matched to a building block database. (B) Comparison of path length and route diversifiability for three FDA-approved drugs: valsartan (**1**), mitapivat (**2**), and pacritinib (**3**). ASKCOS-generated synthetic route for valsartan (**1**) proceeding via (C) a tetrazole ring formation or (D) two alternative Suzuki coupling routes. Building block queries and number of matches in the building block dataset are shown with each building block. (E) Distribution of building block analog similarity to the original building block for molecules that matched the queries from the routes in (D). Example molecules are shown for each set of matches with their similarity to the original building block indicated on the x axis.

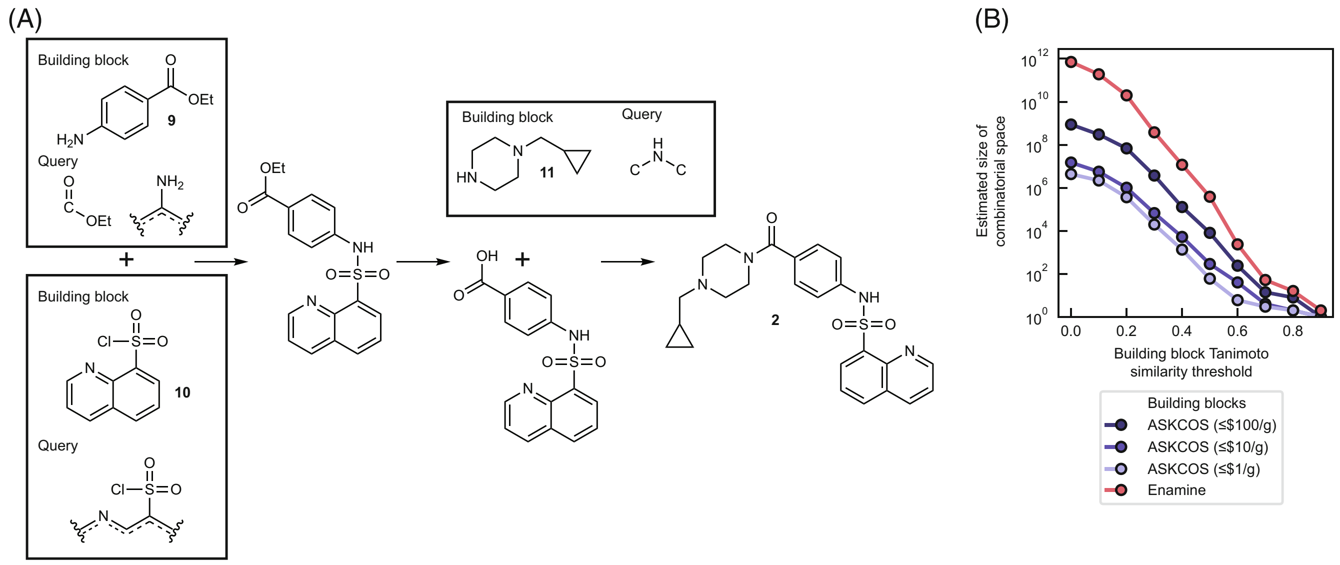

FIGURE 2 Exploring the influence of building block set and price filters on the size of the combinatorial space for an experimentally verified synthesis route. (A) Reported synthesis route for mitapivat (**2**) from Sizemore et al.[33] (B) Estimated size of product analog space for the experimental route given different libraries of building blocks with increasingly stringent similarity filters on the building blocks. Note the log scale used for the y axis.

number of analogs of **1**, we should choose the route that proceeds via a Suzuki coupling with the bromobenzaldehyde building block.

It is important to highlight that the nature of the enumerated chemical analog spaces will be different and not necessarily overlapping between the routes. Following from the reaction sequence, all molecules enumerated based on the route in Figure 1C will have a tetrazole ring in the product but will not necessarily have the biaryl ring structure. Conversely, all of the molecules enumerated based on the routes in Figure 1D will have the biaryl ring structure but will not necessarily have a tetrazole ring. Even between the two alternative Suzuki coupling strategies, the chemical spaces will be different because of the different distributions of matched building block (Figure 1E). Matching the building block analogs to a route before any enumeration provides interpretable insights about the structures of the product analogs, as these follow directly from the structures introduced by the building blocks. The set of matched building blocks is sufficient to understand which modifications from the original product the route enables, and can be used as additional information to help select the synthetic approach.

3 | ANALYZING THE INFLUENCE OF STARTING MATERIAL LIBRARIES ON THE SIZE OF THE ENUMERATED SPACE

The ability to estimate the size of an analog space enables the interrogation of other key parameters in the synthesis planning procedure such as the choice of building block library. For this analysis, we use the experimental synthesis route of mitapivat (**2**) (Figure 2A) from Sizemore et al.[33]

We compared the diversifiability of the route when using the ASKCOS buyables database with a price cutoff of \$100/g (the full database, 2.8×10^5 compounds), \$10/g (1.0×10^5 compounds), and \$1/g (6.1×10^4 compounds) and when using the Enamine make-on-demand building blocks (1.1×10^6 compounds) as the building block library. We find that with the Enamine set and no similarity filter applied to the building blocks, approximately 807 times as many analogs are estimated to be accessible as with the ASKCOS set (Figure 2B). The major contributing factor in this difference is the presence of 16 times as many matches for the secondary amine reactant (**11**) and 13 times as many matches for the sulfonyl chloride (10) in the Enamine dataset. This is a disproportionate increase compared to the overall difference in the database sizes. Approximately four times as many analogs are found for the ester aniline (**9**) building block in the Enamine dataset. As an increasingly stringent Tanimoto similarity filter is applied to the building blocks, the disparity between the estimated size of the enumerated spaces approaches the expected difference in proportion to the size of the two building blocks sets, in this case approximately 4^3. At a threshold similarity of 0.5 between the analog and original building blocks, the estimated difference in the combinatorial space size is a factor of 47 and each of the building block queries match 3–4.5 times more molecules in the Enamine set. As the threshold is raised more, the difference falls further.

The effect of building block price can be interrogated in this manner as well. Setting a maximal price per gram of \$10 and \$1 on the ASKCOS database reduces the number of building blocks from 2.8×10^5 to 1.0×10^5 compounds and 6.1×10^4 compounds, respectively. With a similarity threshold between 0.0 and 0.5 on the building blocks, the size of the enumerated space is approximately 30–60× and 100–200× greater for the full ASKCOS dataset compared to the dataset with the $10 and $1 cutoff, respectively. This kind of sensitivity analysis can help determine whether a route is conducive to diversifiability in more specific discovery contexts where there may be constraints placed on price or building block availability (e.g., in-stock compounds only).

4 | ANALYZING THE INFLUENCE OF BUILDING BLOCK PROPERTIES ON ENUMERATED SPACE PROPERTIES

Size of the analog space is not the only important parameter in a molecular discovery setting. It is also important that the molecules in the enumerated space exhibit properties specific to the particular design objectives. Being able to estimate the number of "shots on goal" in an implicitly enumerated space is useful to compare possible synthesis routes. Filtering building blocks to focus an enumerated space can reduce the cost of downstream screening that requires explicit enumeration or experimental validation, so an understanding of how filters will impact the size and properties of the enumerated space is valuable.

While some molecular properties are complex, nonlinear functions of the component substructures (e.g. clearance), others are more linearly related to the substructures and atoms that make up the molecule. This is obviously true for the molecular weight of a molecule, but is also true for other properties commonly used to approximate "drug-likeness"[34-36] that (a) can be counted, like the number of rotatable bonds or hydrogen-bond donors/acceptors, or (b) can be accurately estimated using group-contribution methods such as topological polar surface area (TPSA)[37] and calculated octanol–water partition coefficient (logP).[38]

We empirically validated that the sum of the molecular weight, TPSA, and calculated logP of the building blocks with a constant correction value for a given synthesis route approximates the computed values for the product molecule. For a random sample of 1 million molecules enumerated using the experimental synthesis of mitapivat (**2**) with the ASKCOS set of buyables, we found a mean absolute difference between the summed quantity and the computed quantity for the enumerated product of 8×10^{-14} g/mol, 5×10^{-3} Å^2, and 3×10^{-15} for molecular weight, TPSA, and calculated logP, respectively. The Pearson correlation coefficient between the summed and the computed values is > 0.9999 for all three properties for this set of molecules (Figure S1). The only one of these properties for which summation did not yield a perfect estimate is TPSA, which is estimated within floating point error for 999,550 of the compounds. For the remaining 450 compounds, there was a consistent error of 10.14Å^2 that can be explained by the discrepancy between the change in TPSA caused by the alkylation of aziridines and non-aziridine secondary amines.[37]

Because these properties are exactly additive (other than floating point errors) or nearly additive with respect to the building blocks once leaving groups are accounted for, the distribution of the product analog properties is well approximated by the distribution of the sums of the building block analog properties. Distributions of sums of independent random variables can be computed by convolving the distributions of the individual random variables. More formally, if a product molecule property (V^P) is additive with respect to the properties of the building blocks (e.g., V^{B_1} and V^{B_2}) with a constant pathway-dependent correction factor for that property C_V that takes into account leaving group effects:

$$V^P = V^{B_1} + V^{B_2} + C_V. \tag{1}$$

It follows that the probability density function of the property, f_{V^P}, is the convolution of the probability density functions of the building blocks, $f_{V^{B_1}}$ and $f_{V^{B_2}}$, giving:

$$f_{V^P}(y - C_V) = (f_{V^{B_1}} * f_{V^{B_2}})(y) = \int_{-\infty}^{\infty} f_{V^{B_1}}(x) f_{V^{B_2}}(y - x)dx. \tag{2}$$

This can be extended to arbitrarily many building blocks as:

$$f_{V^{P'}}(y - C_V) = (f_{V^{B_1}} * f_{V^{B_2}} * \ldots * f_{V^{B_n}})(y). \tag{3}$$

Computing the convolution of distributions using the NumPy Python package is negligible in cost compared to enumerating even a subset of the possible analog space. Furthermore it requires knowledge of the building block properties alone which would need to be calculated one time for a given building block dataset.

Using the experimental synthetic route of **2** with the ASKCOS buyables database, we find that the total size of the possible analog space contains 875,986,622 molecules (607 analogs for **9**, 71 analogs for **10**, and 20,326 analogs for **11**). In addition to the time required to enumerate the product analogs, just instantiating RDKit molecule objects and computing molecular weight, TPSA, and logP for a subset of 68 million analogs took >100 CPU hours compared to approximately 10 CPU seconds to compute those properties for the building blocks and perform the convolution.

Similar to the calculation of diversifiability, the efficiency of the convolution makes it practical as a method to compute additional parameters for use in comparing proposed synthesis plans. It allows rapid estimation of the number of product analogs that fall within desired ranges of molecular weight, lipophilicity, etc. at a speed that safely outpaces the rate with which CASP tools are able to return route suggestions.

Additionally, rapid estimation of the analog space property distributions facilitates further elucidation of the relationship between filters applied to building blocks and the properties of the enumerated analog space. For this analysis, we employed two types of filters: property filters and similarity filters. Property filters were set to constrain physicochemical properties computed for each building block analog. Similarity filters were set to constrain the Tanimoto similarity computed between the 2048-bit Morgan fingerprint representation of the analog and original building blocks. As in Synthesia and Pathfinder, we set uniform filters on all building blocks, so all building analogs for all starting materials that do not pass that filter are removed.

Using the experimentally reported synthesis of mitapivat (**2**) as an example again, we studied the relationship between the distribution

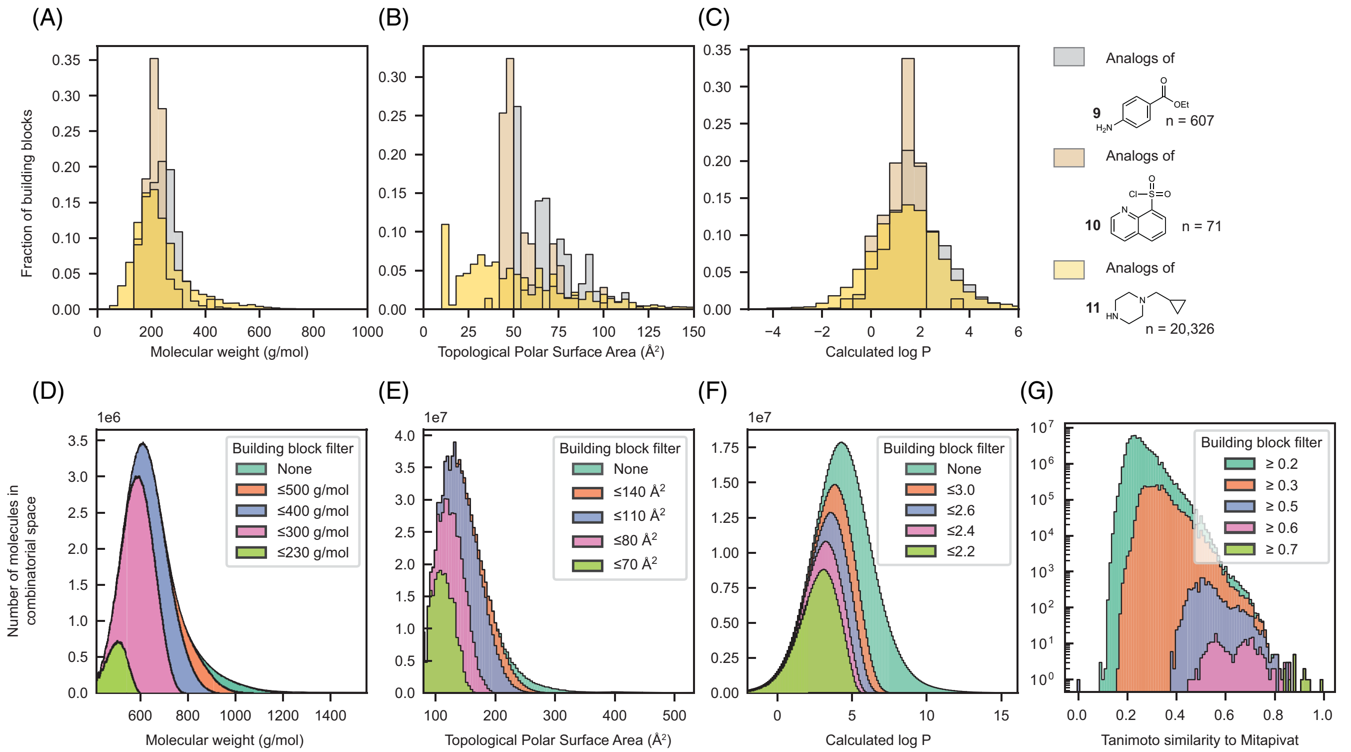

FIGURE 3 (A–C) Property distributions for analogs of the building blocks from the synthesis route of mitapivat (**2**) shown in Figure 2. (D–F) Property distributions for analogs of **2** computed by convolving the property distributions of the building block analogs. Distributions are shown with different filters applied to building blocks. (G) Distribution of product analog similarities to **2** with different thresholds for the similarity between building block analogs and the corresponding original building block.

of building block and product MW, TPSA, and calculated logP (Figure 3A–F). Setting increasingly stringent filters reduces the size of the chemical space by biasing the enumerated chemical space toward desirable property space. Taking the distribution of TPSA as an illustrative example (Figure 3e), approximately 4.4×10^8 of 8.8×10^8 molecules (50%) are estimated to fall below the cutoff of 140Å^2 proposed by Veber et al. [35] for the unfiltered analog space. Setting a cutoff filter on the building blocks of 80Å^2 reduces the absolute number of molecules with a TPSA ≤ 140 Å^2 to 3.7×10^8, but increases the fraction of the enumerated space that satisfies the property constraint to 78%. Even though the focused analog space is smaller in size, it is more enriched for desirable properties and still does not require explicit enumeration, leading to significant speed advantages.

Our property distribution estimations allow us to quantify the tradeoff inherent to setting uniform building block property filters. In addition to uniform building block filters, we can devise a more efficient way to impose property filters on the implicitly enumerated space. We describe this on-the-fly application of filters for additive properties of building blocks that takes into account the contributions from each analog in the following section.

We also studied the distribution of the structural similarities of the analogs to the original product (Figure 3G), as this quantifies how broad of a search is being performed around the original product. This evaluation required explicit enumeration of the products because the Tanimoto similarity of the final product analogs is not additive with respect to the building block similarities. We can see in the overlapping histograms that as building block analogs are constrained to be more similar to the original building blocks, we focus the enumerated chemical space around compounds with high structural similarity to the original product.

5 | APPLICATION OF MULTISTEP REACTION-BASED ENUMERATION TO HIT EXPANSION

To illustrate how we envision integrating synthesis planning and molecular discovery, we applied our computational pipeline to the identification of molecules that score highly according to an "oracle" model for c-Jun N-terminal kinase 3 (JNK3) inhibition.[39,40] Our automated workflow follows five modular stages: a preliminary screen of a diverse chemical dataset to identify the top "hit" molecules, synthesis planning with ASKCOS, route selection based on predicted diversifiability, combinatorial enumeration of analogs, and finally, a screen of the analog space to identify novel hits (Figure 4A).

The oracle model is a random forest classifier trained on experimental bioactivity data from the ExCAPE-DB database.[39,41] Molecules are input as ECFP6 fingerprints.[42] The model output score represents the model's confidence that the molecule is active as an inhibitor against JNK3. We choose this model as it can be applied in a high throughput manner to approximate experimental measures of bioactivity, but the workflow is compatible with any quantitative scoring approach.

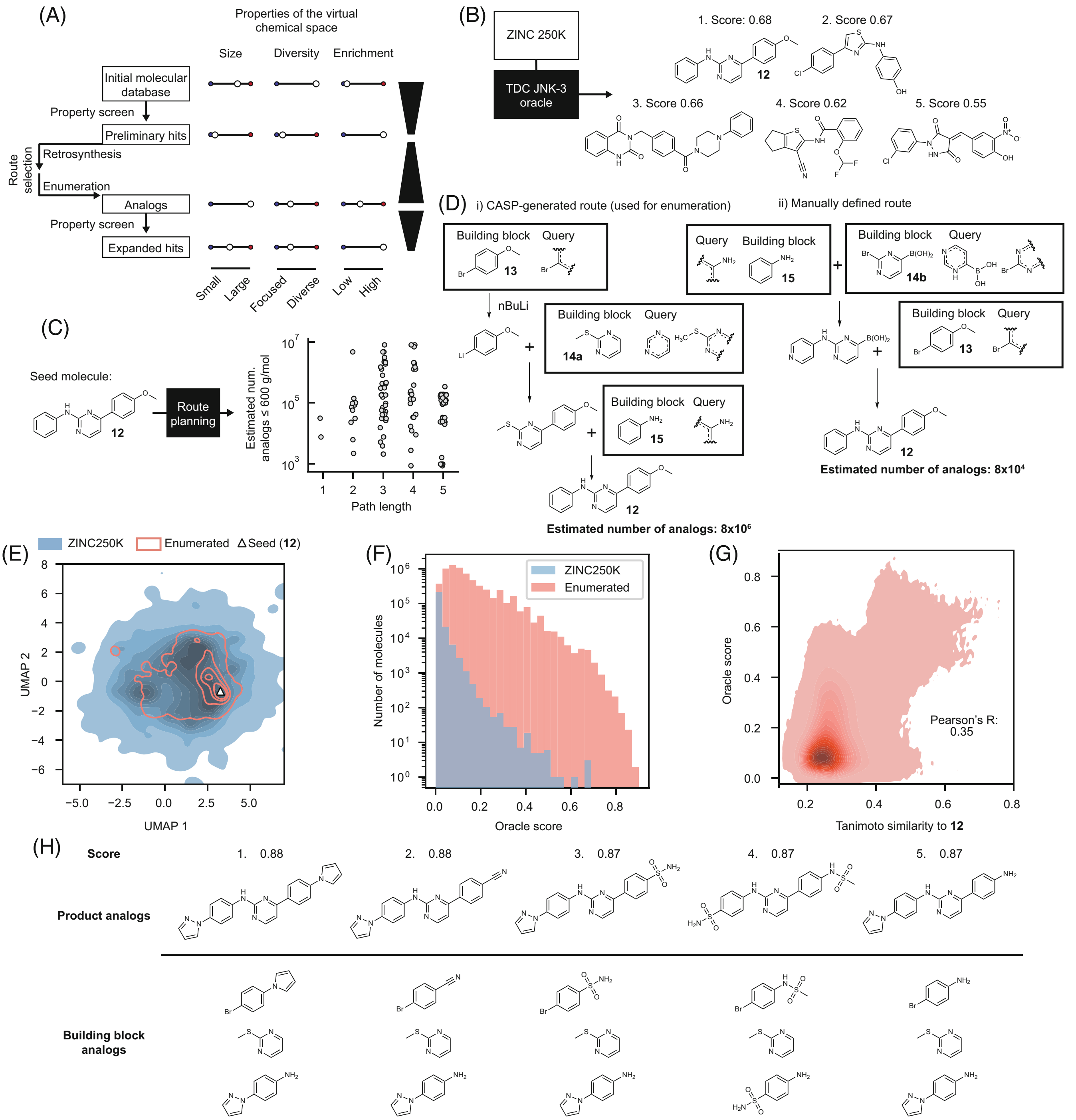

FIGURE 4 (A) Schematic of the back-and-forth discovery pipeline with corresponding properties of the chemical space at each step. (B) Top-scored molecules from the ZINC 250K dataset using the JNK3 activity oracle model. (C) Summary of the routes ASKCOS returned for input molecule **12** with a building block similarity threshold of 0.3 and a product molecular weight cutoff of 600 g/mol. (D) (i) Most diversifiable route returned from ASKCOS and (ii) a manually defined route for **12**. Estimated number of analogs are indicated for the routes with the same Tanimoto similarity threshold and molecular weight cutoff. (E) Overlay of the molecule densities in a two-dimensional representation of chemical space for a random subset of 50,000 molecules from the ZINC250K and 50,000 molecules from the enumerated analogs of **12**. (F) Distributions of predicted JNK3 activity scores for the ZINC250K and enumerated datasets. (G) Density plot of the enumerated chemical space comparing Tanimoto similarity to **12** computed on 2048-bit Morgan fingerprints and predicted JNK3 activity scores. Darker areas indicate a higher density of molecules. (H) Five of the highest-scoring enumerated molecules. Below each product analog is the set of building block analogs that were used to generate it.

We began by screening a dataset of 249,455 drug-like molecules randomly sampled from the ZINC database (ZINC250k[43]). We evaluated each molecule using the JNK3 oracle model and identified the top hits by score (Figure 4B). We elected to proceed with only the top hit (**12**), though the approach could be easily extended to arbitrarily many candidates.

We queried the ASKCOS tree-builder tool for synthesis routes to **12**. A total of 144 plans were returned. Paths that contained regioselectivity concerns when run in the forward direction were automatically flagged and removed, leaving 130 routes. The total number of accessible analogs with a molecular weight ≤ 600 g/mol from each route was estimated using our convolution-based approach with different similarity thresholds on the building blocks ranging from 0 to 0.9; building blocks were further filtered based on substructures that would lead to problematic features in the product using Brenk filters.[44] This analysis led us to select a similarity threshold of 0.3 which corresponds to an enumerated product space of $\leq 8 \times 10^6$ molecules (Figure 4C), which corresponded to the approximate size cutoff we had set for this project based on the available compute power and the time and memory required to enumerate and score the molecules with the oracle model.

The route predicted to yield the greatest number of analogs satisfying the molecular weight cutoff was selected as the input for the product analog enumeration (Figure 4D(i)). This route proceeds through a relatively uncommon, though not unprecedented, coupling between aniline (**15**) and a derivatized 2-methylthiopyrimidine (**14a**).[45,46] We compared the estimated diversifiability of this computer-generated route with a manually defined synthetic route that proceeds via a nucleophilic aromatic substitution followed by a Suzuki coupling (Figure 4D(ii)), with the same product molecular weight and building block Tanimoto similarity filters applied to both routes. We find that the route proposed by ASKCOS is estimated to lead to a chemical space approximately 100 times greater than the chemical space from the manually defined route. Given that two of the building blocks (**13** and **15**) are identical between the routes, the result can be attributed to the fact that there are 84 matches in the building block database for **14a** and none for **14b** with a Tanimoto similarity ≥ 0.3 to the respective molecule. It follows that the enumerated space from the computer-generated route is a strict superset of the space that can be enumerated from the manually defined route, so we proceed with the computer generated route. Whether the larger chemical space or a smaller more chemically tractable chemical space is preferred will ultimately depend on the constraints of a specific discovery project, and can be decided by the user.

The enumeration was performed with "on-the-fly" filtering in addition to a similarity threshold of 0.3. As combinations of building blocks were selected for enumeration, common additive drug-likeness properties (TPSA, number of rotatable bonds, MW, number of hydrogen-bond donors, number of hydrogen-bond acceptors, calculated logP) were computed for each building block and summed, adding a correction factor to account for leaving groups. This approximates properties of the product analog at negligible computational cost prior to enumeration. Cutoffs were set to be slightly more permissive than the criteria for oral bioavailability proposed by Lipinski et al.[34] and Veber et al.[35]: TPSA ≤150, number of rotatable bonds ≤10, number of hydrogen acceptors ≤11, number of hydrogen donors ≤6, and logP ≤6. All combinations that were predicted to satisfy cutoffs set for each property were explicitly enumerated, yielding 6,488,129 compounds. This approach ensures that all enumerated molecules fall within the desired ranges for the target properties. Such filters could only be applied for properties that are additive with respect to building blocks, precluding their application to properties such as structural similarity to the original product. On-the-fly filtering was made efficient by caching computed properties for each building block to avoid redundant calculations.

To gain a qualitative understanding of how the enumerated chemical space compared to the original chemical space of the ZINC250K dataset, we projected a random subset of 50,000 molecules from each set (Figure 4E) into two dimensions. Each molecule was encoded as a 2048-bit Morgan fingerprint and embedded by fitting a UMAP[47] model to the molecules of the ZINC250K dataset and applying it to both sets of molecules. The distribution of enumerated chemical analogs clusters around the seed molecule used to perform the enumeration in chemical space. This supports the premise that reaction-based enumeration enables a local search around a molecule.

Further, the enumerated space is significantly enriched for molecules scored more highly by the oracle model compared to the ZINC250K space (Figure 4F; note the log scale), supporting the premise that a local search is useful in optimizing a candidate hit molecule. The enumeration produced a total of 4377 compounds with scores greater than that of the seed molecule. The top-scored molecules show relatively subtle modifications from the original product molecule (Figure 4G, Figure S2). Molecules that are too similar to the original product do not show an improved score, and molecules that are too dissimilar to the original product likewise do not show an improved score; empirically, we find that the compounds in the enumerated set that achieve the best scores have Tanimoto similarities between 0.4 and 0.6.

Five of the highest scoring molecules are shown in Figure 4H. Of the 4377 compounds with improved scores, all but three were assembled using the original methylthiopyrimidine (**14a**) building block, despite the fact that there were 84 analogs for this building block in the full set of enumerated molecules. The 4-(1H-pyrazol-1-yl)aniline building block was used as an analog of aniline (**15**) to generate 316 of the improved compounds, including 9 out of the top 10. These patterns of enrichment aid in elucidating an interpretable structure activity relationship where the subsequent optimization could potentially be performed in the space of building blocks.

6 | LIMITATIONS

The molecular discovery workflow presented in this article illustrates how implicit enumeration can inform the selection of synthetic routes for an explicit enumeration. While we rely on the number of accessible molecules as a metric to select synthetic routes, larger analog spaces do not necessarily contain a larger number of higher-scoring molecules. Route diversifiability serves as a useful parameter to rank potential synthetic routes in the absence of additional information, but incorporating more complex functions of building block properties will improve the process of route selection.

Further, we acknowledge that the synthetic routes and the reaction templates used to estimate route diversifiability and perform combinatorial enumerations are imperfect. Relying on algorithmically defined reaction templates to assess reaction feasibility likely overestimates route diversifiability. Defining more expressive reaction templates that also describe incompatible molecular patterns to avoid in queried building blocks could mitigate this issue. However, determining the appropriate amount of chemical context to include in reaction templates would necessitate additional reaction data for the proposed transformation or manual curation,[48] limiting the number of routes that could be evaluated. With well-curated reaction templates, for robust 1- and 2-step chemical couplings, Enamine has reported an experimental success rate of approximately 80%[3] for the synthesis of combinatorially generated compounds. We expect that this represents the upper limit on the fraction of a diverse enumerated chemical space that is truly synthesizable. Incorporating data-driven reaction prediction models remains a promising future direction for further promoting the synthesizability of the enumerated space.

7 | CONCLUSION

Synthetic strategies can have downstream effects in influencing the chemical space explored in a molecular discovery project. In this article, we introduced computational tools to rapidly estimate the size ("diversifiability") and property distributions of the chemical space that is synthetically tractable given a synthetic route and set of building blocks with no explicit enumeration of molecules. We demonstrated how these analyses can help rank proposed synthetic routes as well as quantify the impacts of applying building block filters or changing building block libraries. We incorporated the tools into a hit-expansion workflow to select a synthetic route for route-based enumeration and inform the application of filters to focus the enumerated space. Using a surrogate model for bioactivity to score candidate molecules, we identified over 4000 molecules hypothesized to be accessible using the same synthetic route with improved scores compared to the input hit molecule. The workflows developed herein are available as open source code to encourage their application to experimental molecular optimization workflows where synthetic pathway selection and analog design should be coupled.

AUTHOR CONTRIBUTIONS

Itai Levin: conceptualization (equal); methodology (equal); software (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Michael E. Fortunato:** conceptualization (equal); methodology (equal); software (equal); writing – review and editing (equal). **Kian L. Tan:** validation (equal); writing – review and editing (equal). **Connor W. Coley:** conceptualization (equal); project administration (equal); supervision (equal); writing – review and editing (equal).

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R21AI169342. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Itai Levin and Connor W. Coley thank the Machine Learning for Pharmaceutical Discovery and Synthesis consortium for additional support.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The numerical data for the plots in all of the figures are available as Appendix S1. Appendix S1 also includes the synthetic routes generated by ASKCOS for the target molecules **1**, **2**, **3**, and **12** used for the subsequent analyses, the manually defined synthetic route for mitapivat, and the list of molecule SMILES from the ZINC250K and enumerated compound sets and their scores predicted with the JNK3 activity oracle model. Additionally, the code for the workflows presented in this article is available at https://github.com/itai-levin/easie. The repository includes scripts to perform analog count estimation or explicit analog enumeration for a set of routes automatically generated by ASKCOS or for a manually defined synthetic route. The scripts can take building block or product property filters as input to constrain the counting or enumeration, and allow users to specify the path name for the set of building blocks. No original experimental data is associated with the article.

ORCID

Itai Levin https://orcid.org/0000-0001-8881-8162
Michael E. Fortunato https://orcid.org/0000-0003-1344-5642
Kian L. Tan https://orcid.org/0000-0001-8243-1223
Connor W. Coley https://orcid.org/0000-0002-8271-8723

REFERENCES

1. Shen Y, Borowski JE, Hardy MA, Sarpong R, Doyle AG, Cernak T. Automation and computer-assisted planning for chemical synthesis. *Nat Rev Method Primer*. 2021;1(1):1-23. doi:10.1038/s43586-021-00022-5
2. Hu Q, Peng Z, Sutton SC, et al. Pfizer global virtual library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb Sci*. 2012;14(11):579-589. doi:10.1021/co300096q
3. Grygorenko OO, Radchenko DS, Dziuba I, Chuprina A, Gubina KE, Moroz YS. Generating multibillion chemical space of readily accessible screening compounds. *iScience*. 2020;23(11):101681. doi:10.1016/j.isci.2020.101681
4. Irwin JJ, Tang KG, Young J, et al. ZINC20 - a free ultra large-scale chemical database for ligand discovery. *J Chem Inf Model*. 2020;60(12):6065-6073. doi:10.1021/acs.jcim.0c00675
5. Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of Ultralarge compound collections for drug discovery. *J Chem Inf Model*. 2022;62(9):2021-2034. doi:10.1021/acs.jcim.2c00224
6. Walters WP. Virtual chemical libraries. *J Med Chem*. 2019;62(3):1116-1124. doi:10.1021/acs.jmedchem.8b01048
7. Kaplan AL, Confair DN, Kim K, et al. Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity. *Nature*. 2022;610(7932):582-591. doi:10.1038/s41586-022-05258-z
8. Gorgulla C, Boeszoermenyi A, Wang ZF, et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*. 2020;580(7805):663-668. doi:10.1038/s41586-020-2117-z
9. Fink EA, Xu J, Hübner H, et al. Structure-based discovery of nonopioid analgesics acting through the α2A-adrenergic receptor. *Science*. 2022;377(6614):eabn7065. doi:10.1126/science.abn7065

10. Luttens A, Gullberg H, Abdurakhmanov E, et al. Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J Am Chem Soc*. 2022;144(7):2905-2920. doi:10.1021/jacs.1c08402
11. Zwick CR, Sosa MB, Renata H. Modular chemoenzymatic synthesis of GE81112 B1 and related analogues enables elucidation of its key pharmacophores. *J Am Chem Soc*. 2021;143(3):1673-1679. doi:10.1021/jacs.0c13424
12. Yang Z, Shi S, Fu L, Lu A, Hou T, Cao D. Matched molecular pair analysis in drug discovery: methods and recent applications. *J Med Chem*. 2023;66(7):4361-4377. doi:10.1021/acs.jmedchem.2c01787
13. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci*. 1998;38(3):511-522. doi:10.1021/ci970429i
14. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*. 2008;3(10):1503-1507. doi:10.1002/cmdc.200800178
15. Brown N, McKay B, Gilardoni F, Gasteiger J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J Chem Inf Comput Sci*. 2004;44(3):1079-1087. doi:10.1021/ci034290p
16. Firth NC, Atrash B, Brown N, Blagg J. MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation. *J Chem Inf Model*. 2015;55(6):1169-1180. doi:10.1021/acs.jcim.5b00073
17. Polishchuk P. CReM: chemically reasonable mutations framework for structure generation. *J Chem*. 2020;12(1):28. doi:10.1186/s13321-020-00431-w
18. Vinkers HM, de Jonge MR, Daeyaert FFD, et al. SYNOPSIS: SYNthesize and OPtimize system in silico. *J Med Chem*. 2003;46(13):2765-2773. doi:10.1021/jm030809x
19. Hartenfeller M, Zettl H, Walter M, et al. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol*. 2012;8(2):e1002380. doi:10.1371/journal.pcbi.1002380
20. Pophale R, Daeyaert F, Deem MW. Computational prediction of chemically synthesizable organic structure directing agents for zeolites. *J Mater Chem A*. 2013;1(23):6750-6760. doi:10.1039/C3TA10626H
21. Spiegel JO, Durrant JD. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J Chem*. 2020;12(1):25. doi:10.1186/s13321-020-00429-4
22. Bradshaw J, Paige B, Kusner MJ, Segler MHS, Hernández-Lobato JM. Barking up the right tree: an approach to search over molecule synthesis DAGs. *arXiv:201211522 [cs, q-bio]* 2020.
23. Gao W, Mercado R, Coley CW. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. *arXiv:2110.06389 [cs.LG]* 2021.
24. Konze KD, Bos PH, Dahlgren MK, et al. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J Chem Inf Model*. 2019;59(9):3782-3793. doi:10.1021/acs.jcim.9b00367
25. Bos PH, Houang EM, Ranalli F, et al. AutoDesigner, a de novo design algorithm for rapidly exploring large chemical space for Lead optimization: application to the design and synthesis of d-amino acid oxidase inhibitors. *J Chem Inf Model*. 2022;62(8):1905-1915. doi:10.1021/acs.jcim.2c00072
26. Dolfus U, Briem H, Rarey M. Synthesis-aware generation of structural analogues. *J Chem Inf Model*. 2022;62(15):3565-3576. doi:10.1021/acs.jcim.2c00246
27. Genheden S, Bjerrum E. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. *Dig Dis*. 2022;1:527-539. doi:10.1039/D2DD00015F
28. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. 2018;555(7698):604-610. doi:10.1038/nature25978
29. Coley CW, Thomas DA, Lummiss JAM, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*. 2019;365(6453):eaax1566. doi:10.1126/science.aax1566
30. Schwaller P, Hoover B, Reymond JL, Strobelt H, Laino T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci Adv*. 2021;7(15):eabe4166. doi:10.1126/sciadv.abe4166
31. Coley CW, Green WH, Jensen KF. RDChiral: an RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J Chem Inf Model*. 2019;59(6):2529-2537. doi:10.1021/acs.jcim.9b00286
32. Himo F, Demko ZP, Noodleman L, Sharpless KB. Mechanisms of tetrazole formation by addition of azide to nitriles. *J Am Chem Soc*. 2002;124(41):12210-12216. doi:10.1021/ja0206644
33. Sizemore JP, Guo L, Mirmehrabi M, Su Y. Crystalline forms of n-(4-(4-(cyclopropylmethyl) piperazine-1-carbonyl)phenyl)quinoline-8-sulfonamide. WO2019104134A1. 2019.
34. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 2001;46(1-3):3-26. doi:10.1016/s0169-409x(00)00129-0
35. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the Oral bioavailability of drug candidates. *J Med Chem*. 2002;45(12):2615-2623. doi:10.1021/jm020017n
36. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of Druglike properties. *ACS Chem Nerosci*. 2010;1(6):435-449. doi:10.1021/cn100008c
37. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem*. 2000;43(20):3714-3717. doi:10.1021/jm000942e
38. Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci*. 1999;39(5):868-873. doi:10.1021/ci990307l
39. Li Y, Zhang L, Liu Z. Multi-objective de novo drug design with conditional graph generative model. *J Chem*. 2018;10(1):33. doi:10.1186/s13321-018-0287-6
40. Huang K, Fu T, Gao W, et al. Artificial intelligence foundation for therapeutic science. *Nat Chem Biol*. 2022;18(10):1033-1036. doi:10.1038/s41589-022-01131-2
41. Sun J, Jeliazkova N, Chupakhin V, et al. ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *J Chem*. 2017;9(1):17. doi:10.1186/s13321-017-0203-5
42. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742-754. doi:10.1021/ci100050t
43. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci*. 2018;4(2):268-276. doi:10.1021/acscentsci.7b00572
44. Brenk R, Schipani A, James D, et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*. 2008;3(3):435-444. doi:10.1002/cmdc.200700139
45. Itami K, Yamazaki D, Ji Y. Pyrimidine-core extended π-systems: general synthesis and interesting fluorescent properties. *J Am Chem Soc*. 2004;126(47):15396-15397. doi:10.1021/ja044923w

46. Byth KF, Culshaw JD, Green S, Oakes SE, Thomas AP. Imidazo [1,2-a]pyridines. Part 2: SAR and optimisation of a potent and selective class of cyclin-dependent kinase inhibitors. *Bioorg Med Chem Lett*. 2004;14(9):2245-2248. doi:10.1016/j.bmcl.2004.02.015
47. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426 [stat.ML]* 2018.
48. Heid E, Goldman S, Sankaranarayanan K, Coley CW, Flamm C, Green WH. EHreact: extended Hasse diagrams for the extraction and scoring of enzymatic reaction templates. *J Chem Inf Model*. 2021;61(10):4949-4961. doi:10.1021/acs.jcim.1c00921

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Levin I, Fortunato ME, Tan KL, Coley CW. Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways. *AIChE J*. 2023;69(12):e18234. doi:10.1002/aic.18234