JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 252−261

pubs.acs.org/jcim

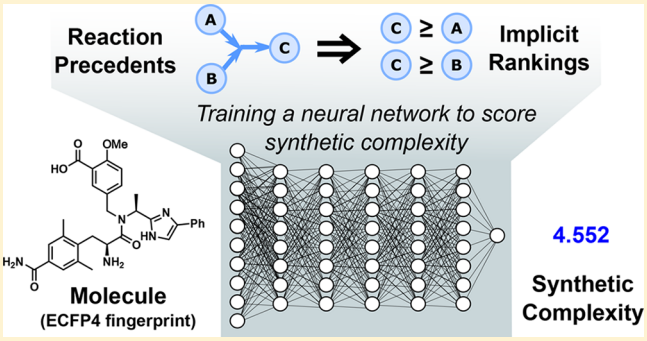
SCScore: Synthetic Complexity Learned from a Reaction Corpus

Connor W. Coley, Luke Rogers, William H. Green,* and Klavs F. Jensen*

Department of Chemical Engineering, Massachusetts Institute of Technology; 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: Several definitions of molecular complexity exist to facilitate prioritization of lead compounds, to identify diversity-inducing and complexifying reactions, and to guide retrosynthetic searches. In this work, we focus on synthetic complexity and reformalize its definition to correlate with the expected number of reaction steps required to produce a target molecule, with implicit knowledge about what compounds are reasonable starting materials. We train a neural network model on 12 million reactions from the Reaxys database to impose a pairwise inequality constraint enforcing the premise of this definition: that on average, the products of published chemical reactions should be more synthetically complex than their corresponding reactants. The learned metric (SCScore) exhibits highly desirable nonlinear behavior, particularly in recognizing increases in synthetic complexity throughout a number of linear synthetic routes.



INTRODUCTION

The difficulty of quantifying "complexity" is apparent when simply considering the numerous ways one can define complexity—should a large multifunctional, multichiral compound be considered complex? While the total synthesis of a steroid is long and challenging, a synthesis starting from readily available compounds like cholesterol might only require a few reaction steps.

Herein, we focus on the specific definition of complexity that is "synthetic complexity". In simple terms, we want molecules that are easy to synthesize to have a low score, while molecules that are hard to synthesize to have a high score. Synthetic complexity—or the opposite, synthetic accessibility—can be used in drug discovery (as well summarized by Méndez-Lucio and Medina-Franco[1]) as a scoring metric for virtual compound libraries and to assist medicinal chemists in prioritizing certain lead compounds over others,[2,3] for designing compound libraries in diversity-oriented synthesis,[4] or for computer-assisted synthesis planning as a metric for determining how promising each candidate disconnection is while performing a retrosynthetic search.[5−7]

The standard procedure for quantifying complexity (including both synthetic complexity and its more general counterpart, "molecular complexity") is to either (a) construct a heuristic definition based on domain expertise or (b) design a model that can be regressed on ground truth data. Expert chemists may be asked to score drug-like molecules to validate an expert-defined heuristic in the former case or to fit a regression in the latter case. Existing approaches to quantifying complexity can be broadly categorized into the following:

(1) Scores calculated from a small number of expert-selected structural descriptors[8−12]
(2) Scores calculated through some sort of automated retrosynthetic analysis[13−16]
(3) Scores calculated by regressing a large number of structural descriptors[17]
(4) Scores calculated using fragment or substructure-level contributions[19−24]

All of these techniques must be regressed, trained, or validated on some ground truth data. For retrosynthetic techniques, this may be the ease of finding a synthetic route starting from simpler starting materials, sometimes as a binary easy/hard score.[16] For numerical scores, often a survey of expert chemists is used as the data set.[12,17,18] Some studies have examined what features are most prevalent in large chemical databases, most notably Ertl and Schuffenhauer's use of the PubChem database to identify rare structural motifs when defining their SA_Score metric[21] or Fukunishi et al.'s[22] use of commercially available compound databases to correlate complexity with price.

Despite numerous studies seeking to quantify complexity, published reactions are rarely used as the basis for scoring. One exception is the recently-developed metric from Li and Eastgate, the "current complexity", that quantifies a molecule's complexity based partially on the complexity of a known synthetic route to that molecule; this allows for a complexity score to change over time as more concise syntheses are developed, but requires the use of retrosynthetic planning software to generalize to new compounds.[18] Another important exception is the approach

Received: October 23, 2017
Published: January 8, 2018

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

proposed by Heifets,[25] who trains a model to predict the number of synthetic steps needed to reach a given chemical from buyable compounds based on 43,976 molecules from the patent literature. Each molecule was labeled with the number of reaction steps required to produce it from a fixed set of commercially available compounds.

We propose a methodology for using precedent reaction knowledge to learn a function for evaluating synthetic complexity. This synthetic complexity score provides an additional metric by which virtual screening or molecular design pipelines can prioritize compounds as a complement to other metrics. Specifically, we train a neural network model to quantify synthetic complexity scores in a manner that correlates with the number of reaction steps required to produce a compound, but does not rely on the availability of multistep reaction pathways, chemist rankings, or rigid encoding of buyable compounds for training. We refer to this learned complexity score as the SCScore.

APPROACH

Desirable Properties. Our definition of synthetic complexity arises from one simple premise: *if a synthetic route between compounds A and C goes through intermediate B, then the reaction steps A → B and B → C should each be considered productive reaction steps, reflected by an increase in synthetic complexity.* There are, of course, many domain-specific definitions of what a "reasonable" or "good" synthetic route is. However, a synthetic complexity score should trend upward during a multistep synthesis.[8,25] Equivalently, the retrosynthetic analysis of a target compound should allow us to find a pathway to buyable chemicals with only "downhill" moves that reduce complexity. This enables the prioritization of retrosynthetic disconnections during an automated retrosynthesis search. This prioritization has traditionally been done with simple expert heuristics like the length of molecules' canonical SMILES[26] representation raised to the 3/2 power ($SMILES^{3/2}$).[5] More generally, aligned with the intent of one of the earliest complexity metrics from Bertz,[8] the metric should reflect when progress is being made in a synthetic sequence.

We would also like the synthetic complexity score to be insensitive to the exact starting materials database. For a specific retrosynthetic program, it might make sense to specify precisely which compounds are available. For a universal scoring function, we prefer to keep this definition general and work under the assumption that the most synthetically simple compounds are reasonable starting materials, either purchased directly or synthesized straightforwardly. Chemicals commonly used as reactants are necessarily easier to obtain than products, which provides an indication of what might be commercially available without restricting the analysis to depend on a specific compound database.

Details of Approach. We aim to find a suitable function f with flexible parametrization (θ) to map any molecule m to a numerical synthetic complexity score (eq 1).

$$\text{SCScore}(m) \equiv f(m;\, \theta) \tag{1}$$

In order to satisfy the premise of our definition of synthetic complexity, we seek to impose an inequality constraint on the complexity assigned to chemicals appearing in each reaction example $\{R_1 + R_2 + ... + R_n \rightarrow P\}$, where R_i are the reactants and P is the major product. As shown in eq 2, this means that the score assigned to a reaction product should be at least as large as any of the scores assigned to the corresponding reactants.

$$f(P;\, \theta) \geq \max\{f(R_i;\, \theta)\}_i \;\; \forall \; (R_1 + R_2 + ... + R_n \rightarrow P) \tag{2}$$

To formalize this training objective, we divide our set of reaction examples $\{R_1 + R_2 + ... + R_n \rightarrow P\}$ into a set of molecular pairs {R, P} for which the constraint should hold. Rather than seek an exact solution to achieve this separation, we define a hinge loss function to match this pairwise ranking objective

$$L(\theta) = \sum_{(R,P)} \max(0, f_0 - [f(P;\, \theta) - f(R;\, \theta)]) \tag{3}$$

where the inclusion of $f_0 > 0$ encourages a minimum degree of separation between reactants and products. $L(\theta)$ represents a sum of "penalties" from each reactant and product pair and is the function that we aim to minimize during model training. Note that because the model $f(m; \theta)$ takes a single molecule as input and outputs a single score, two model evaluations are used per (R, P) pair to calculate the difference in their scores.

An important consequence of using a hinge loss function (i.e., $\max(0, x)$) instead of a binary cross-entropy loss function (as would be more typical for a pairwise ranking task) is that the model will not try to maximize the separation of scores between reactants and products. If the model recognizes a reactant as significantly simpler than a product—with a difference exceeding f_0—then that example does not contribute to $L(\theta)$ and will not be used to update model parameters. This mitigates the risk of overfitting, where molecules that only appear as products would be inclined toward the maximum score; such molecules need to be only slightly more complex than the reactants required to synthesize them. During training, a large fraction of examples will readily satisfy $f(P;\, \theta) - f(R;\, \theta) \geq f_0$, while a much smaller number of examples will contribute to the model's learned nuance. This is analogous to how a support vector regression model may depend on a subset of training examples as support vectors, although all examples have the potential to affect the final model.

A visualization of this approach is shown in Figure 1. Starting from a database or network of known reactions (Figure 1a), we wish to find a means of arranging all known chemicals so that products are scored more highly than each of their corresponding reactants (Figure 1b) as a means of assigning each a quantitative SCScore (Figure 1c). If we do so using an appropriately parametrized model, the scoring function will be generalizable to novel compounds.

Our reaction-based definition affords the following advantages:

(1) The perceived difficulty of synthesizing a molecule is directly informed by historical trends pertaining to how products are made and—implicitly—in how many steps. Each molecule is analyzed in the context of all known molecules as they appear in all known reactions.

(2) The SCScore is an aggregate of experimental information and does not suffer from potential personal biases (e.g, reactions very familiar to chemists on paper might actually be used with low frequency in practice) or from the limitations of automated retrosynthetic analyses (i.e., missing reaction rules or making unrealistic disconnections). Previous studies have shown significant disagreement between chemists when scoring synthetic accessibility and when prioritizing compounds during drug discovery, suggesting that a data-driven model could provide a useful complement to subjective chemist evaluations.[27,28]

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

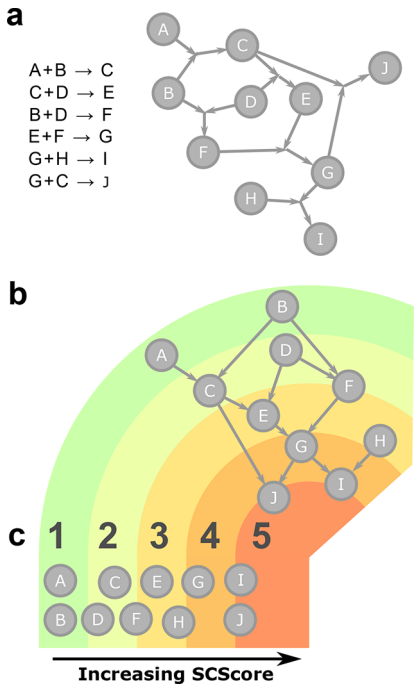


Figure 1. Visualization of the approach to quantifying synthetic complexity (SCScore). Starting from a reaction network of known reactions (a), we wish to find a means of arranging all known chemicals so that products are scored more highly than each of their corresponding reactants (b) as a means of assigning a quantitative synthetic complexity score to every known compound (c) using a parametrized model that is generalizable to novel compounds. In reality, the network of known reactions used for training and testing consists of millions of examples involving millions of distinct chemical species.

(3) By reformulating the problem of predicting synthetic complexity as an analysis of reactions instead of substances, we avoid the need for ground truth (molecule, value) pairs and instead learn from ranked (molecule, molecule) pairs. This enables the use of much larger data sets (i.e., reaction databases) for training instead of small, expert-defined data sets.

(4) Synthetic complexity should be nonlinear with respect to the number of atoms or functional groups present; it might be that it is easier to prepare the protected R-NHBoc version of a compound rather than the R-NH_2, which should be reflected by the primary amine having a higher synthetic complexity score. This precludes the use of simple additive models combining fragment contributions, as the whole molecule must be analyzed simultaneously to understand this nuance. It also necessitates the analysis of reactions to understand when protections or deprotections should be considered productive steps.

Implementation. We use a Morgan fingerprint of radius 2 (similar to an Extended-Connectivity Fingerprint of diameter 4 (ECFP4)[29]) as implemented in the RDKit,[30] including chirality, folded to a length 1024 bit vector as the input to our model. Two other fingerprint representations, length-2048 boolean fingerprints and length-1024 integer fingerprints, were also tested but were not found to offer significant advantages (Tables S4 and S5). The model, implemented in Python 2.7 using Tensorflow,[31] is a feed forward neural network consisting of five hidden layers with 300 hidden nodes each, ReLU activation with bias, and a single sigmoid output. No regularization (e.g., L2, Dropout) was used. The output score $\in (0, 1)$ is linearly scaled to $(1, 5)$ to be more natural for human interpretation. We set the offset in the hinge loss to be $f_0 = 0.25$ so that an "equilibrium" separation of all intermediates in a 16-step synthesis could fit within the range $(1, 5)$. Because synthetic complexity—as defined—should be nonlinear with respect to structure, we did not explore simpler model architectures that may have enabled straightforward interpretation as a fragment-contribution approach does. A schematic of the model and more training details can be found in the Supporting Information (Figure S1).

As our data source, we use the ca. 12 million reactions in Reaxys[32] that have (a) a reaction structure file parseable by RDKit and (b) at least one set of reaction details that is labeled as a "single-step" reaction and (c) have a single recorded major product. This last criterion ensures that trivial side products (e.g., salts, water) are excluded. These are expanded to ca. 22 million (reactant, product) pairs. We use a randomized 80:10:10 training:validation:testing split. Note that although entries are not filtered by publication year, there is a natural bias toward recently published reactions because the rate of publishing is growing exponentially.

Assumptions of Approach. In how we have structured the learning task, a molecule is considered "hard to make" (i.e., would be assigned a high score) if it appears to require many reaction steps to synthesize through conventional means. There are some subtleties in this assumption that warrant mention:

(1) Although the objective during optimization is to impose a separation between reactants and products of at least $f_0 = 0.25$, some reactions will be more complexifying than others and not all reactions will achieve this separation. In practice, it might be possible to synthesize a compound with score 3.5 in fewer steps than one with score 3.0 if 3.5 appears to be synthetically challenging but is actually conducive to useful synthetic steps or a convergent synthesis. The more a compound appears as a reactant, the lower its perceived synthetic complexity will tend to be.

(2) There is no explicit knowledge of what starting materials are available (i.e., purchasable). This is learned implicitly by the model as it gradually learns what types of structures tend to be present in reactants and what structures tend to be present in products. If a certain motif appears in both the reactants and products of a single reaction example, then a gradient step trying to increase the separation of their scores will not change how that motif contributes to the overall SCScore. Conceptually, the learned SCScore thus mimics how a chemist might identify a certain substructure as something that can be purchased and installed, rather than created. This can result in large discrepancies between the SCScore and past synthetic accessibility scores when complex scaffolds are perceived to be readily available due to the prevalence of similar compounds as reactants. As an example, Deoxycholic acid is perceived to be a "low complexity molecule" because it appears significantly more often at the start of a synthetic step than at the end; a linear route to Deoxycholic acid is shown later in Figure 7.

(3) There is no explicit consideration of how complicated a particular reaction is in terms of its necessary reagents, catalysts, procedural difficulty, or even its yield. This makes the learned synthetic complexity score more suitable for application to a research setting (e.g., a medicinal chemistry laboratory) rather than a develop-

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

ment setting because process chemistry necessitates stricter considerations of what constitutes a good synthesis.[21,33] These considerations are correspondingly less able to be captured by a single metric. In Sheridan et al.'s study, this mismatch manifested itself as a relatively poor correlation between estimated synthetic complexity and process mass intensity (PMI) compared to its correlation with SA_Score.[17] No attempt was made in this study to categorize reactions from the Reaxys database into research-relevant and process-relevant subsets.

(4) The SCScore model will be biased by the types of reactants and products that appear in the data (Reaxys). This means that complicated natural products (of which there are relatively few in Reaxys) may "saturate" our scoring function and be assigned a maximum value of 5. We see this with several examples from Sheridan et al. later in Figure S3, where chemicals are assigned a value close to 5 (corresponding to the maximum complexity).

RESULTS

Quantitative Evaluation. Most reactant−product reaction pairs are easily separated during training, as indicated by the steep loss curves shown in Figure 2 and the rapid increase in the fraction of examples with a separation greater than $f_0 = 0.25$. Validation performance was used to select the best performing model state, which was then evaluated using the test set and subjected to further qualitative analysis. The similar performance between the training data and the validation data suggests that the model is not being significantly overfit, which is consistent with the comparable test set performance achieved after final model selection. Additional scatterplots showing correlation between training/validation losses and validation/testing losses are shown in Figure S2.

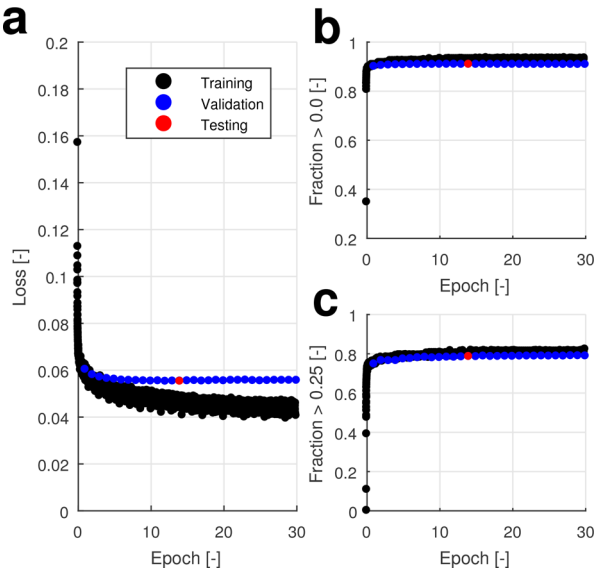


Figure 2. Profiles showing the (a) average hinge loss, (b) fraction of reactant−product pairs with a positive SCScore difference, and (c) fraction of reactant−product pairs with a SCScore difference exceeding 0.25. Each epoch represents one iteration of training on the entire data set. The validation performance (blue) was used to determine the model state to use for testing (red) and further application.

There are 2,214,928 reactant−product pairs in the testing set. Learned complexity scores are shown in aggregate for these compounds in Figure 3. Here, 91.3% of these pairs have a positive SCScore difference and 78.9% have an SCScore difference greater than $f_0 = 0.25$. The median increase in SCScore is 0.77, although it is important to keep in mind that these statistics are for all reactant−product pairs. A significant number of reactants are trivial salt fragments or counterions that are assigned a low SCScore very close to 1 (as shown in Figure 3a).

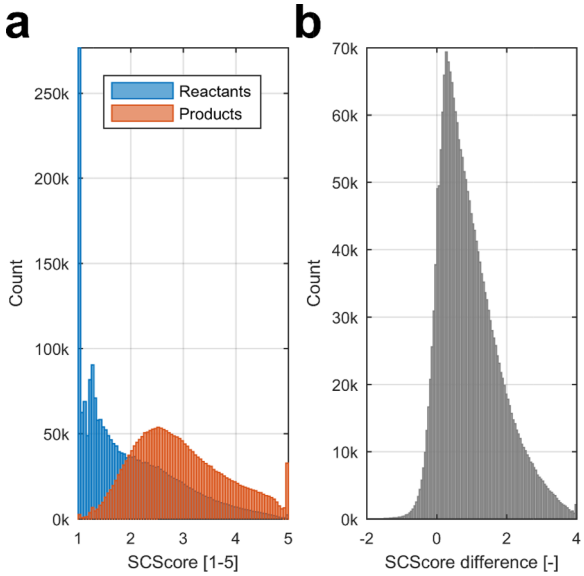


Figure 3. Histograms showing (a) the SCScore assigned to each reactant and product in the test set and (b) the difference in SCScores for each reactant−product pair. Many reactant fragments are trivial salts or counterions and are assigned low scores near 1.

Visualization of Synthetic Complexity. To visualize the distribution of assigned synthetic complexity scores as a function of molecular structure, principal component analysis was applied to the fingerprints of a random selection of 50,000 products appearing in the test set. The first three principal components are used to produce the plot shown in Figure 4a, where each point is colored by its SCScore (red: high; blue: low). Because the learned model uses the molecular fingerprint as its only input, we expect a correlation to exist. However, the trend observed globally does not hold upon closer examination of a narrow region (−0.1 < PCA3 < 0.1), shown in Figure 4b. The same visualizations using the SMILES$^{3/2}$ heuristic function are shown in Figure 4c and d; the SMILES$^{3/2}$ score exhibits a similar global trend but does not exhibit the same degree of local variation. This is explained by the neural network's ability to capture effects that are nonlinear with respect to the original fingerprint. Not only does the model have the capacity to capture nonlinear effects, but Figure 4 shows that this nonlinearity is necessary to capture the nuances of synthetic complexity, as it is defined here, beyond what is reflected by the first three principal components.

Comparison to Chemist Scores. Sheridan et al.[17] provide a large open-source set of structures (divided into several subsets) and complexity scores assigned by a large panel of expert chemists. Out of 1775 compounds, 44 cannot be parsed by the most recent release of RDKit[30] and were excluded from the comparison. We assign complexity scores to the remaining 1731 compounds using our learned SCScore model as well as the Synthetic Accessibility score (SA_Score);[21] these are shown in Figures S3 and S4, respectively. The full data set, including the 44 unparseable molecules, can be found in the Supporting

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

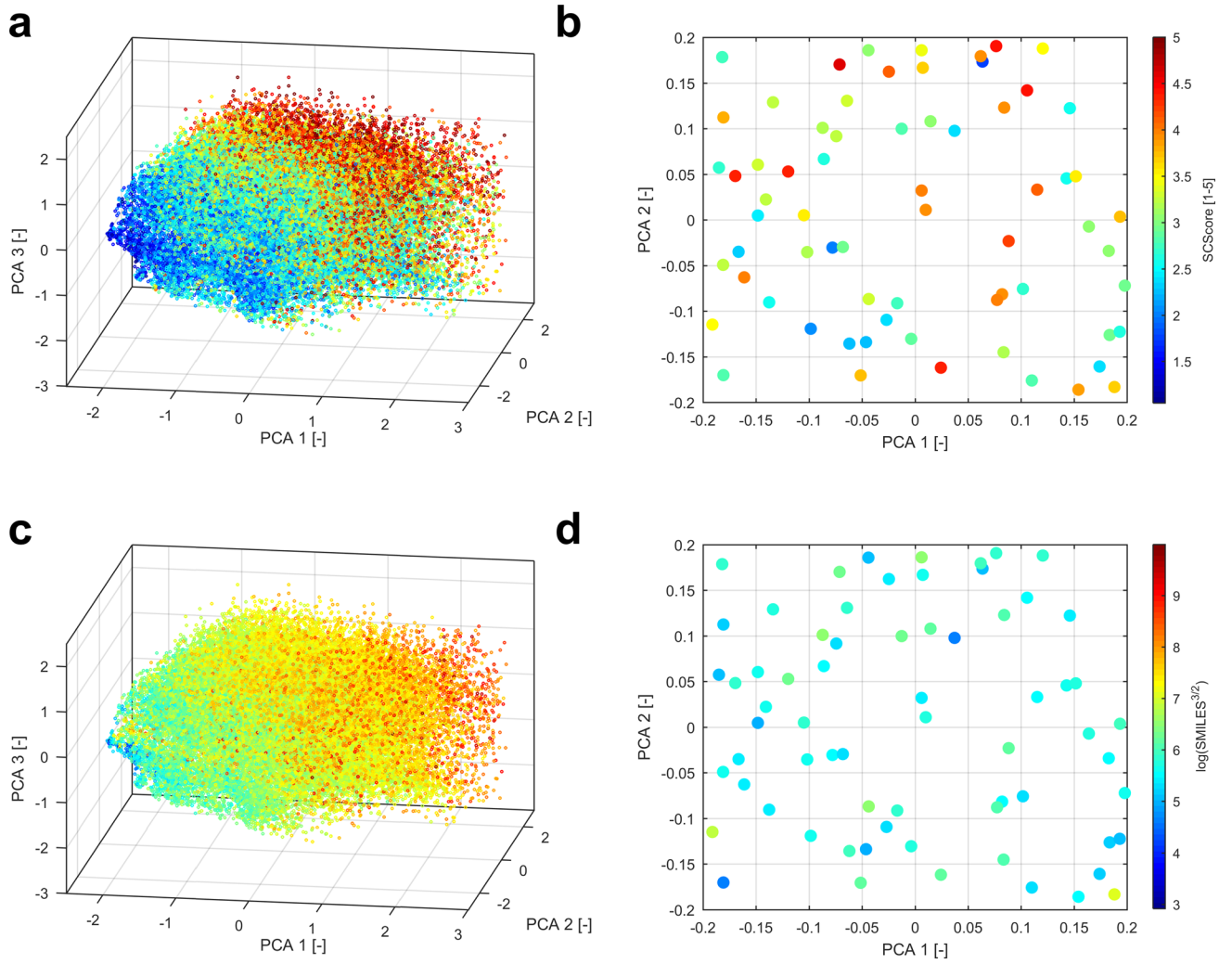


Figure 4. Visualization of complexity for 50,000 random products from the Reaxys test set. (a, c) Scatterplot of the first three principal components, colored by assigned complexity score. (b, d) Zoomed-in view of a slice with −0.1 < PCA3 < 0.1 for (a, b) the learned SCScore and (c, d) the heuristic SMILES$^{3/2}$ score, where in all cases red signifies a higher score than blue. Both the SCScore and heuristic SMILES score show a similar global trend in complexity as a function of the first three principal components. However, the SCScore shows a high degree of local variation due to the nonlinearity of the trained model, while the heuristic SMILES$^{3/2}$ score shows very little variation as this simple scoring heuristic does not capture any nuance in synthetic complexity.

Information. Several molecules from Sheridan et al.s MDDR (MDL Drug Data Report) subset are substantially more complex than those most commonly found in Reaxys, which results in a "saturation" of the SCScore at the maximum complexity of 5. Histograms showing the distributions of scores can be found in Figure 5, demonstrating that our model consistently ranks molecules as more synthetically complex than Sheridan et al.; this is particularly true for the MDDR subset (Figure S6).

A statistically significant correlation exists between Sheridan et al.'s meanComplexity score and the SCScore assigned here for all data subsets except the small Kjell[33] set, consisting of just 15 molecules and previously used in a study of process mass intensity (PMI); within the Kjell et al. subset, the correlation to meanComplexity is stronger using SA_Score. This is not entirely surprising, as the compounds used for training the SA_Score and the compounds used in Sheridan et al. are both intended to be drug like. Interestingly, even the naive SMILES $^{3/2}$ heuristic complexity score correlates strongly with the chemist consensus meanComplexity, arguably more closely than both the SCScore and SA_Score (Figure S5).

The lackluster correlation with Sheridan et al.'s mean-Complexity scores should not be taken as an indication of a poor model for synthetic complexity. In their crowdsourcing study, "explicit instructions stated that the goal was to score "complexity" and not specifically synthetic difficulty."[17] The intent of the study was to analyze an aggregated definition of molecular complexity based on subjective chemist scores, rather than to study one specific definition of complexity. Similarly, the SA_Score model from Ertl and Schuffenhauer[21] was developed as a complexity metric based on the popularity of fragments in the PubChem database, which is correlated with but perhaps not directly aligned with synthetic complexity. The disparity between meanComplexity scores and SCScores is related to the SCScore's focus on synthetic complexity, rather than the broader definition of molecular complexity that was captured by Sheridan et al.'s survey.

Several compounds with exceptionally large discrepancies are shown in Figure 6. The first compound (Figure 6a), glucosamine, is perceived by the model as synthetically simple due to its prevalence as a starting material, which implies that it is

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

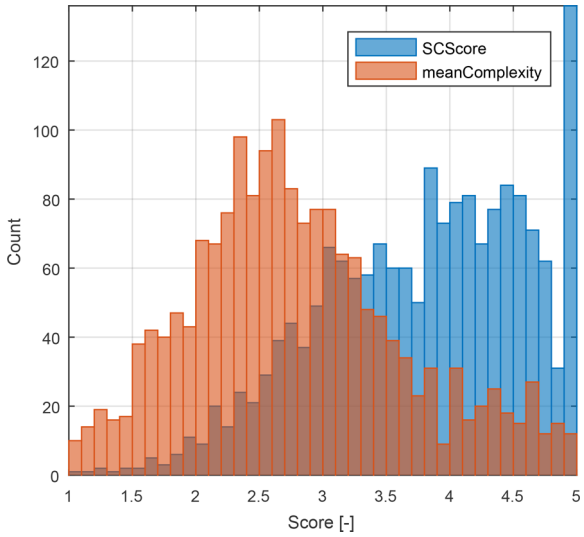


Figure 5. Distributions of the SCScore and the meanComplexity score as assigned by chemists in Sheridan et al.[17] for the entire data set. Several molecules from the MDDR subset are substantially more complex than what is most commonly found in Reaxys, which results in a "saturation" of the learned score at the maximum complexity of 5.

straightforward to obtain (in this case, to purchase); the chemist meanComplexity score was 3.105 ± 1.314, the largest standard deviation of the entire data set, suggesting that some survey respondents recognized it as a common building block while some did not. As Sheridan et al. explain, "a molecule that is complex as seen by one method (e.g., with many chiral centers) may appear very synthetically accessible in a retrosynthetic view if most of the chiral centers are contained in a single preexisting reagent".[17] Similarly, 1,2,3,4,5,6-cyclohexanol (Figure 6b) is trivial to purchase and also received a wide range of chemist scores (reported standard deviation of 1.26). Ouabain (Figure 6c) is not inexpensive, but is commercially available and is more commonly employed as a reactant than a product. The fourth example (Figure 6d) strongly resembles the penicillin core structure and is perceived as being a straightforward derivative of such an available starting material. The remaining four compounds (Figure 6e–h) are perceived by the model as synthetically complex but by chemist respondents as structurally simple. Compounds in Figure 6e and f both contain stereoselectively substituted tetralin scaffolds, which accounts for their perceived complexity.

Analysis of Synthetic Routes. A major goal for our learned synthetic complexity score is to reflect when progress is being made in a multistep synthesis; this is a result of its design to correlate with the number of synthetic steps required to produce a molecule. To evaluate its performance in this regard, we turn to a recent paper by Flick et al.[34] describing likely synthetic routes to the 29 new chemical entities (NCEs) approved in 2015, divided into 41 linear syntheses. Many steps actually consist of multiple parts; we refer the reader to the original paper for details. We calculate the synthetic complexity of every starting material, intermediate, and final product in these 41 linear syntheses using (a) our SCScore, (b) SA_Score,[21] as implemented in RDKit, and (c) a simple heuristic SMILES$^{3/2}$ based on the length of a compound's SMILES string representation. A comparison of our learned scoring function against SA_Score is shown in Figure 7 as a function of reaction step; a similar plot showing the SMILES$^{3/2}$ score is shown in Figure S7. An ideal trend would be a monotonic increase in synthetic complexity with reaction step number, operating under the assumption that every reaction step in these syntheses helps make progress toward the goal of synthesizing the final product molecule. The SMILES strings and scores for all 205 structures can be found in the Supporting Information.

Qualitatively, SCScores (blue) tend to reflect a monotonic increase in synthetic complexity throughout these linear syntheses more often than the SA_Score (orange), which was designed to measure synthetic complexity but was trained on substances rather than reactions. Selected syntheses are shown in Figure 8.

Perhaps the most striking difference is observed in the synthesis of Polmacoxib (Figure 8a), where the SA_Score perceives a *decrease* in synthetic complexity for four of the six reaction steps. If the products of these reactions were actually more synthetically accessible than their reactants, then the overall synthesis could be trivially improved by starting with the product and forgoing that reaction. The SCScore model perceives a monotonic increase in complexity as the linear synthesis proceeds. Similar trends can be seen for the synthesis of

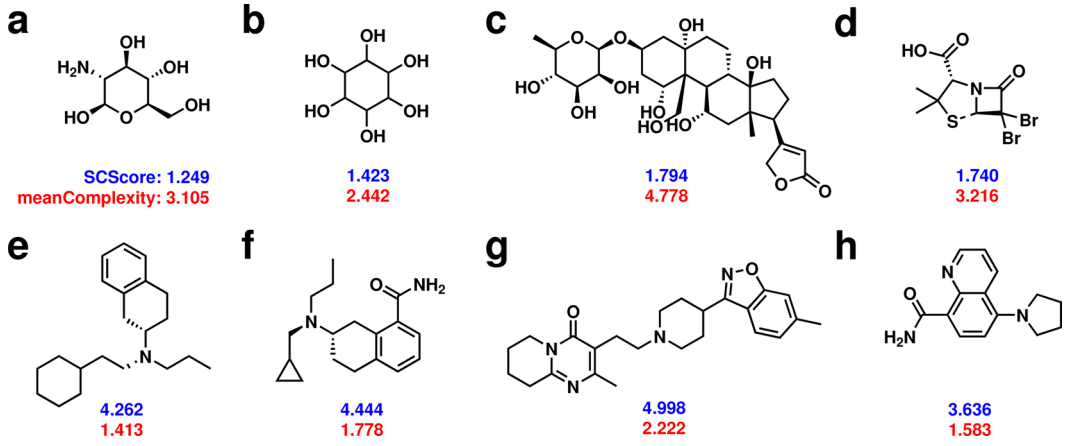


Figure 6. Compounds from Sheridan et al.[17] with substantial differences between the SCScore (blue) and chemist consensus meanComplexity score (red), both between 1 and 5. (a–d) Compounds that the model perceives to be simple that were scored by chemists as complex, perhaps because the chemists did not consider or recognize what is buyable. (e–h) Compounds that the model perceives to be complex that were scored by chemists as simple.

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252–261

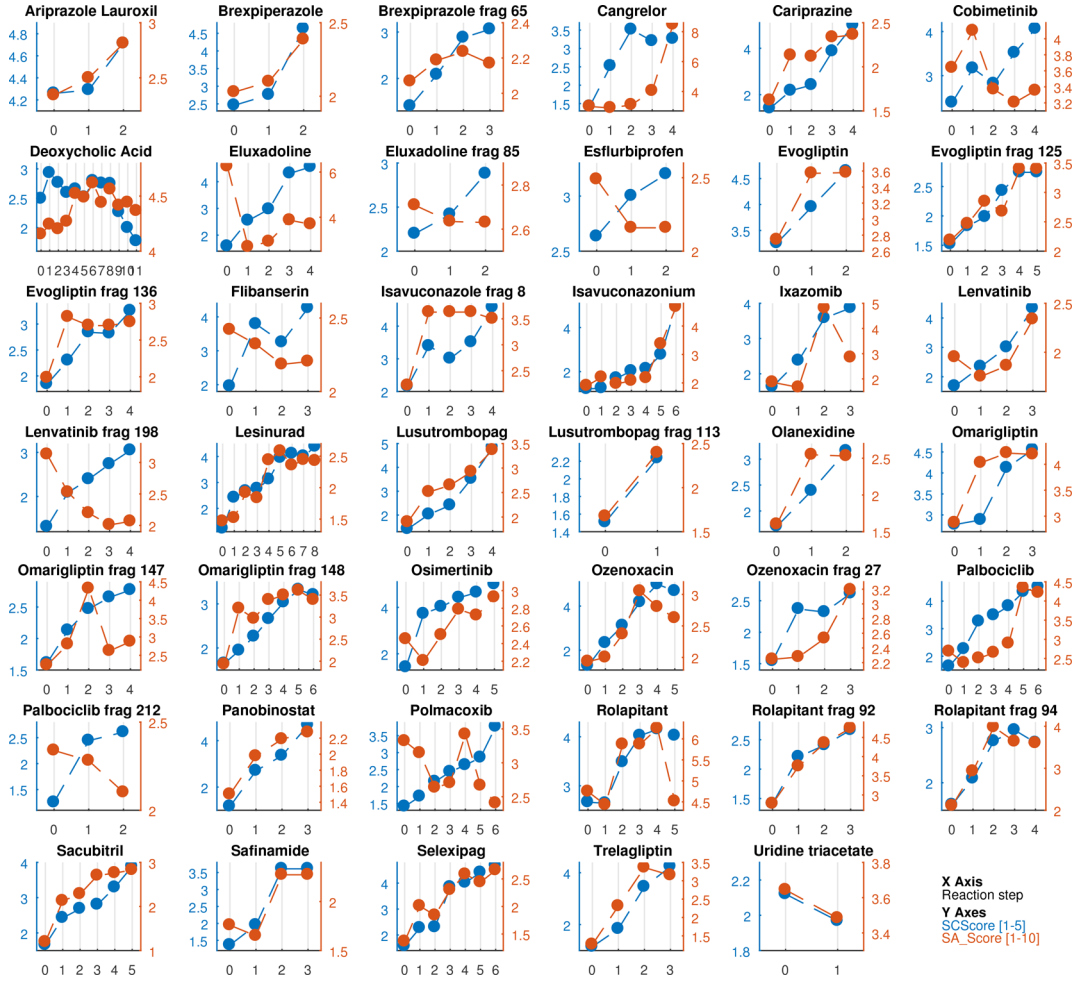


Figure 7. Synthetic complexity scores assigned to each reactant, intermediate, and product in 41 linear syntheses of the 29 new chemical entitites (NCEs) approved in 2015 as reported by Flick et al.[34] SCScores are shown in blue, while SA_Scores are shown in orange. *X*-axes correspond to the reaction step number with starting materials at $x = 0$.

Lenvatinib fragment 198 (Figure 8b), where 3/4 and 0/4 reaction steps are seen as simplifying by SA_Score and our learned model, respectively. There are still several cases in Figure 7 where the SCScore model believes a reaction to be simplifying, one of which can be seen in the second reaction step in the synthesis of Flibanserin (Figure 8d). The removal of the ispropyl group is seen as simplifying; the model does not recognize that substructure as a common protecting group (as it might for, e.g., Boc). The starting material is a convenient source of the central scaffold with broken symmetry and appears to be available from several dozen vendors; the commercial availability of this starting material may have informed the development of the route, which is not captured by the model.

Part of the power of using a flexible neural network model to learn synthetic complexity—rather than relying on heuristic calculations or expert-defined descriptors—is its ability to capture nonlinear or potentially counterintuitive trends. An example of this is shown in Figure 9 for the final step in the synthesis of Eluxadoline. In this reaction step, a methyl ester is hydrolyzed to the acid and a secondary R-NHBoc is deprotected to form the primary R-NH_2 amine. Although this step might reduce the *structural* complexity as perceived by SA_Score, the fact that the synthesis of Eluxadoline relies on this protected intermediate indicates that the intermediate is necessarily less *synthetically* complex. The deprotected Eluxadoline requires strictly more synthetic steps when proceeding through the reported route. The ability of our model to perceive when protections or deprotections are productive is critically important to its potential application to automated retrosynthesis; this retro deprotection step would still be perceived as a "downhill" step, making progress toward reaching simpler, more available starting materials.

Analysis of Reaction Types. To determine the perceived complexity of different reaction types, we turn to the open source ca. 50,000 reaction data set previously for the task of reaction role assignment[35] and retrosynthesis prediction,[36,37] derived from a larger collection from the U.S. patent literature.[38] The reactions of this particular subset have been classified by Schneider et al.[39] into 10 reaction classes described in Table S1. We refer the reader to Schneider et al.[39] for more information on the classification methodology. We use the precleaned data from Coley et al.,[37] whereby examples with multiple products are split into multiple distinct examples. A second set of 14,000 reactions from 14 subclasses within class 1 (Table S1) was taken from Schneider et al.[39] for additional analysis.

Rather than dividing examples into reactant−product pairs, here we define the complexity of a reaction to be the difference in complexity of the product and the maximum reactant complexity. That is,

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

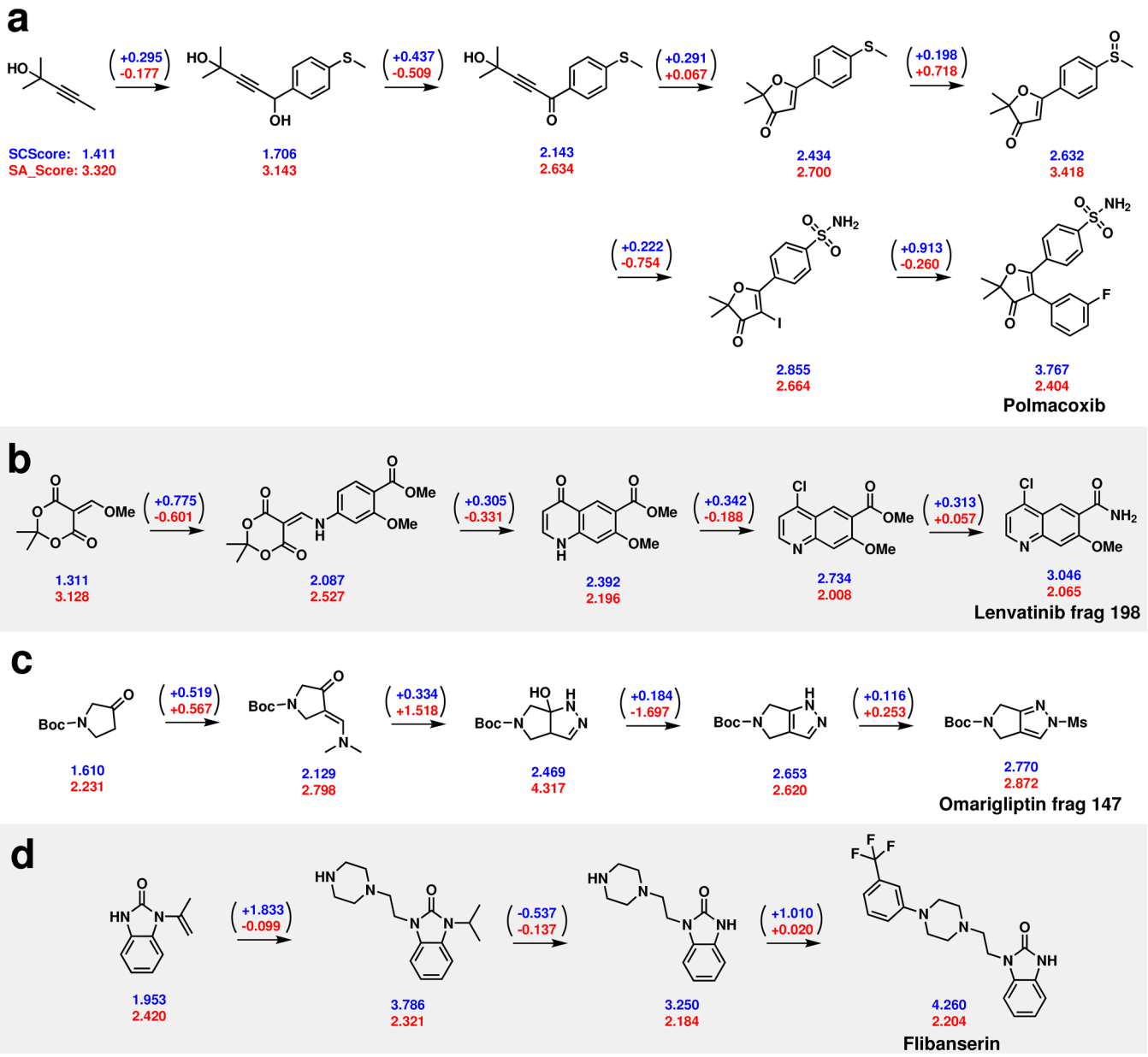


Figure 8. Exemplary syntheses from Flick et al.[34] where there is a significant difference in the complexity trends as evaluated by the learned SCScore model (blue) and the SA_Score (red). The change in complexity is shown above each forward reaction arrow.

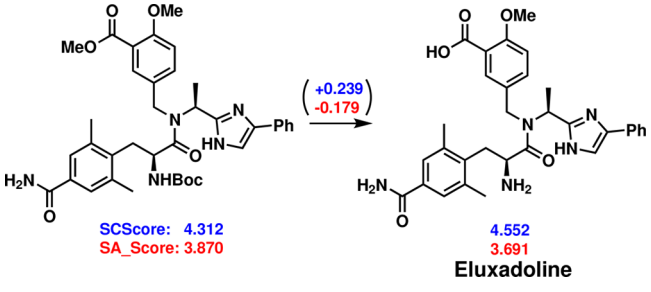


Figure 9. Final step in the synthesis of Eluxadoline,[34] showing scores assigned by the SCScore model (blue) and SA_Score (red). Our model recognizes that the protected form is less synthetically complex, despite the presence of additional functionalities, which is necessarily true due to its use as an intermediate during the preparation of Eluxadoline.

$$f(\mathrm{R}_1 + \ldots + \mathrm{R}_n \rightarrow \mathrm{P}) = f(\mathrm{P}) - \max\{f(\mathrm{R}_i)\} \quad (4)$$

Histograms showing the distributions of reaction complexity assigned to reactions within each class are shown in Figure 10.

Figure 10 reveals that reactions of the first four classes (heteroatom alkylation and arylation, acylation and related processes, C−C bond formation, and heterocycle formation) tend to introduce the most complexity as perceived by the model. This is a consequence of having defined reaction complexity using the maximum of reactant complexities, as this definition favors convergent syntheses where two building blocks of moderate complexity may be brought together to form a product of significantly higher complexity. Reactions within the first four classes tend to be convergent. Heterocycle formation reactions, by definition, form heterocycle motifs that can appear challenging to synthesize in addition to bringing two similarly complex building blocks together, which accounts for their slightly higher average complexity of +0.76; recognizing highly complexifying heterocycle forming reactions is important for

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

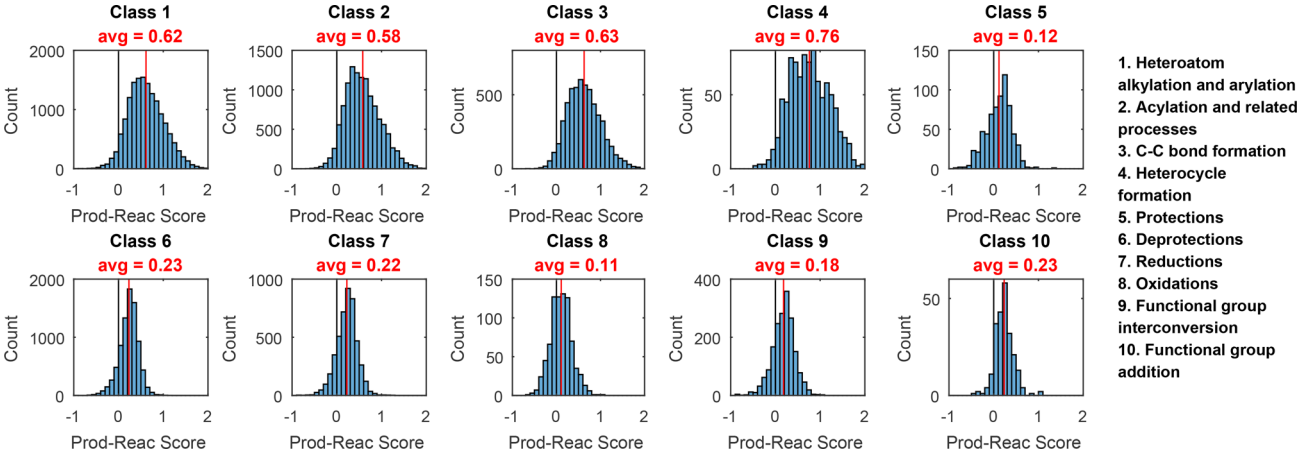


Figure 10. Quantification of the extent to which different reaction types are complexifying as perceived by the SCScore. Histograms depict the differences in scores between products and reactants, where the complexity of multiple reactants is equal to the maximum of the set of reactant complexities. Reaction data is from Schneider et al.[39]

automated retrosynthesis, as these reactions can be overlooked during manual retrosynthetic analysis.

The remaining reaction classes, described in the legend of Figure 10, are more centrally distributed around $\Delta f = 0$, but all have an average increase in synthetic complexity. The fact that both protections and deprotections have positive average SCScore differences is another testament to the model's understanding of when the addition or removal of protecting groups is productive; a fragment-based approach would necessitate that protection and deprotection steps have opposing effects on synthetic complexity. Some types of reactions, particularly protections and oxidations, are largely perceived as "lateral" steps leading to small changes in synthetic complexity. This is consistent with how certain reactions are used in route planning: protections and deprotections are almost always a means to an end, but they themselves may not have a strong effect on synthetic complexity. Likewise, oxidation reactions are often minor functional group manipulations (e.g., alcohol to ketone).

A detailed analysis of subclasses within class 1 (heteroatom alkylation and arylation) can be found in Figure S9 with subclass descriptions in Table S3. An analogous figure to Figure 10 using SA_Score is shown in Figure S8. Seven of the ten reaction classes are perceived to be simplifying (i.e., have a negative average change in complexity), demonstrating the superiority of the SCScore in capturing the increase in synthetic complexity from reactants to products.

CONCLUSION

We have developed a methodology for quantifying synthetic complexity based directly on published reaction data. By reformulating the problem, we shift from the traditional regression methods trained on (molecule, value) pairs to a method that imposes inequality constraints on (molecule, molecule) pairs. The trained SCScore model exhibits justifiable discrepancies with previous definitions of "complexity", namely, the crowdsourced definition from Sheridan et al., as we are specifically focusing on synthetic complexity. SCScores are applicable to lead prioritization in that even if a compound is perceived as complex by expert chemists a low SCScore indicates that there may be an easier way of accessing that compound than initially recognized (Figure 6). SCScores also describe actual multistep syntheses well (Figures 7 and 8), including protections/deprotections (Figures 9 and 10), in a manner that enables its use as a guiding metric for retrosynthesis to identify options for rapidly introducing complexity. We believe that the SCScore has the potential to become an important additional metric in virtual screening pipelines and/or *de novo* molecular design.

We describe our workflow in full detail and open source our code to enable use of other data sets as knowledge bases, for example, in-house electronic lab notebook data. We also include the trained model as a "deployed" model with minimal dependencies for straightforward integration into python-based workflows.

ASSOCIATED CONTENT

Supporting Information

Additional figures and discussion of code and model hyperparameters (code is available at https://github.com/connorcoley/scscore). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00622.

Additional figures and discussion of code and model hyperparameters. (PDF)

Details of comparison to Sheridan et al.'s mean-Complexity. (XLSX)

Details of linear syntheses in Figure 7. (XLSX)

Molecules used for PCA visualization. (TXT)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: whgreen@mit.edu (W. H. Green).

*E-mail: kfjensen@mit.edu (K. F. Jensen).

ORCID

Connor W. Coley: 0000-0002-8271-8723

William H. Green: 0000-0003-2603-9694

Klavs F. Jensen: 0000-0001-7192-580X

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the DARPA Make-It program under Contract ARO W911NF-16-2-0023. C.W.C. received additional funding from the NSF Graduate Research Fellowship Program

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261

under Grant No. 1122374. The authors thank Wengong Jin for help with the code. All code used for model training can be found at https://github.com/connorcoley/scscore.

REFERENCES

(1) Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discovery Today* **2017**, *22*, 120−126.
(2) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369.
(3) Baber, J.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681−692.
(4) Lee, D.; Sello, J. K.; Schreiber, S. L. Pairwise Use of Complexity-Generating Reactions in Diversity-Oriented Organic Synthesis. *Org. Lett.* **2000**, *2*, 709−712.
(5) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(6) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, *33*, 469−476.
(7) Proudfoot, J. R. Molecular Complexity and Retrosynthesis. *J. Org. Chem.* **2017**, *82*, 6968−6971.
(8) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.
(9) Bertz, S. H. Convergence, Molecular Complexity, and Synthetic Analysis. *J. Am. Chem. Soc.* **1982**, *104*, 5801−5803.
(10) Bertz, S. H. On the Complexity of Graphs and Molecules. *Bull. Math. Biol.* **1983**, *45*, 849−855.
(11) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269−272.
(12) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in which Compounds are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.
(13) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2613−2633.
(14) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-Like Molecules. **2011**, *51*, 2768−2777.
(15) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311−325.
(16) Podolyan, Y.; Walters, M. A.; Karypis, G. Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods. *J. Chem. Inf. Model.* **2010**, *50*, 979−991.
(17) Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'Shea, P. D. Modeling a crowdsourced definition of molecular complexity. *J. Chem. Inf. Model.* **2014**, *54*, 1604−1616.
(18) Li, J.; Eastgate, M. D. Current Complexity: a Tool for Assessing the Complexity of Organic Molecules. *Org. Biomol. Chem.* **2015**, *13*, 7164−7176.
(19) Hendrickson, J. B.; Huang, P.; Toczko, A. G. Molecular Complexity: a Simplified Formula Adapted to Individual Atoms. *J. Chem. Inf. Model.* **1987**, *27*, 63−67.
(20) Allu, T. K.; Oprea, T. I. Rapid Evaluation of Synthetic and Molecular Complexity for in silico Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1237−1243.
(21) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8.
(22) Fukunishi, Y.; Kurosawa, T.; Mikami, Y.; Nakamura, H. Prediction of Synthetic Accessibility Based on Commercially Available Compound Databases. *J. Chem. Inf. Model.* **2014**, *54*, 3259−3267.
(23) Bottcher, T. An Additive Definition of Molecular Complexity. *J. Chem. Inf. Model.* **2016**, *56*, 462−470.
(24) Proudfoot, J. R. A Path Based Approach to Assessing Molecular Complexity. *Bioorg. Med. Chem. Lett.* **2017**, *27*, 2014−2017.
(25) Heifets, A. Automated Synthetic Feasibility Assessment: A Data-driven Derivation of Computational tools for Medicinal Chemistry. Thesis, University of Toronto, 2014.
(26) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(27) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47*, 4891−4896.
(28) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the Mind of a Medicinal Chemist: the Role of Human Bias in Compound Prioritization During Drug Discovery. *PLoS One* **2012**, *7*, e48476.
(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(30) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed November 20, 2016).
(31) Abadi, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2016, arXiv:1603.04467.
(32) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society, 2014; pp 127−148.
(33) Kjell, D. P.; Watson, I. A.; Wolfe, C. N.; Spitler, J. T. Complexity-Based Metric for Process Mass Intensity in the Pharmaceutical Industry. *Org. Process Res. Dev.* **2013**, *17*, 169−174.
(34) Flick, A. C.; Ding, H. X.; Leverett, C. A.; Kyne, R. E.; Liu, K. K. C.; Fink, S. J.; O'Donnell, C. J. Synthetic Approaches to the New Drugs Approved During 2015. *J. Med. Chem.* **2017**, *60*, 6480−6515.
(35) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56*, 2336−2346.
(36) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.
(37) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237−1245.
(38) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Thesis, University of Cambridge, 2012.
(39) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39−53.

DOI: 10.1021/acs.jcim.7b00622
J. Chem. Inf. Model. 2018, 58, 252−261