Synthetic route design & assessment using vectors derived from similarity and complexity

Samuel Genheden[a], Gareth P. Howell[b]*

[a] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

[b] Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield SK10 2NA, UK

1. Abstract

With the aim of improving the machine-interpretation of synthetic routes we describe a new theoretical approach to visualising and assessing synthetic pathways in the absence of empirical data such as yield, cost and waste. The representation of molecular structures as coordinates derived from molecular (fingerprint) similarity and complexity allows individual transformations to be viewed as vectors (reactant to product) whereby the magnitude and direction of travel can be used to assess and quantify transformation efficiency. Vectors derived in this way are shown to follow logical trends when grouped by reaction type/class. Synthetic routes can thus be visualised as a series of head-to-tail vectors (one per transformation or step) traversing the range between starting material and target whereby the efficiency with which this range is covered can be quantified. Our approach is built upon the analysis of > 350k literature syntheses (> 1.4m reactions), is readily machine-interpretable and avoids the challenges associated with automated reaction class assignment and atom-mapping.

2. Keywords

3 to 10 keywords

3. List of Abbreviations

CASP – computer aided synthesis planning

ECFP4 – enhanced connectivity fingerprint (diameter 4)

FGA – functional group addition

FGI – functional group interconversion

LLS – longest linear sequence

PMI – process mass intensity

SAR – structure activity relationship

SMILES – simplified molecular-input line-entry system

4. Introduction

The assessment and comparison of synthetic routes in organic chemistry can, after suitable training and experience, be readily accomplished by humans. Regardless of the criteria we are assessing against (cost, time, waste etc), someone "skilled in the art" can make a judgement as to whether the route represents a logical and efficient series of chemical transformations. This is typically achieved by considering the structural complexity of the target and assessing the number and type of transformations in the route, the order in which they are carried out and any reliance on protecting groups, auxiliaries etc.

To do the same assessment on hundreds, thousands or millions of synthetic routes, our suitably-trained chemist quickly becomes the rate-limiting step. If empirical information (e.g., yield or waste) is available, then computerisation is trivial. If such empirical information is either unavailable or unreliable, the task becomes significantly more challenging due to the sparsity of meaningful or generally-accepted metrics.

Step count – either longest linear sequence (LLS) or total – is by far the most common gauge against which synthetic routes are assessed. It is easy to conceptualise, machine-interpretable and a reasonable predictor of the quantitative metrics we are ultimately interested in. If defined and counted consistently, it is also an excellent tool for comparing synthetic routes from a specified starting material to a specified target – with fewer steps usually being better. On its own, step count is less useful for describing or comparing routes where the starting material and/or the target vary. Whilst a human might compare, say, a three step route to one target with a six step route to another and give an assessment which is "better", this is a much more challenging task for a machine. Further limitations with step-count arise when considering literature routes. It is clear that step counting should stop when the desired target structure is reached, but there is no accepted convention for when to begin. Typically, step-counting begins at the first material (working backwards) that can be purchased, regardless of cost or availability. Alternatively, counting begins at

the first material whose synthesis has been reported and deemed “simple”. These approaches are practical, since exhaustively step-counting back to hydrocarbons and biomass feedstock is unrealistic. The result is a high degree of inconsistency, however, with unspecified steps upstream of the starting materials being unaccounted for. Although less common, complications can also arise due to the lack of convention when defining “one step”. Sequential transformations carried out in one vessel (with or without intermediate work-up and/or solvent change) may be reported as a single or multiple steps.

Machines are capable of interpreting transformations and assigning them to a predefined class using commercial software such as NameRxn [1] or InfoChem [2]. This can be valuable since certain classes, for example redox manipulations or functional group interconversions (FGI) can be penalised in favour of “constructive” steps where bonds present in the target skeleton are formed. This strategy remains imperfect since the classification of reactions is prone to failure, particularly when considering novel or tandem transformations. Furthermore, the binary assignment of transformations as productive or non-productive is somewhat limiting for the purposes of comparison or ranking.

Other metrics, such as atom economy[3], step economy [4,5,6], redox economy [7], ideality [8] and convergence [9] have been reported to assess aspects of efficiency relating to synthetic routes. These concepts are all eminently logical and automatable, provided fully atom-mapped synthetic sequences (including reagents, this can be far from accurate)[10]; none are widely used or reported, however, when assessing or reporting synthetic routes.

With the above in mind, we have an interest in novel, automatable strategies for the assessment of synthetic routes that do not rely on empirical data (yield etc) and circumvent the need for complete atom mapping or reaction class assignment.

5. Results and discussion

1. Dataset compilation and shape

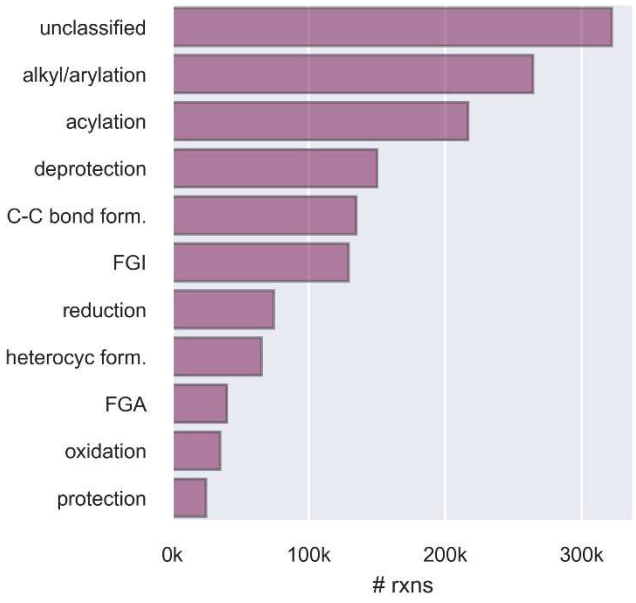
To begin our investigation we compiled a dataset of approximately 350,000 synthetic routes from the period 2000 to 2020. As sources, *Angewandte Chemie International Edition* (ACIE), *The European Journal of Medicinal Chemistry* (EJMC), *The Journal of Medicinal Chemistry* (JMC) and *Organic Process Research and Development* (OPRD) were used. From this time period and these journals, reactions available in Reaxys [11] were initially grouped by citation. Then, per citation, reaction networks were constructed and routes extracted using depth first search as described previously for patent data [12, 13]. In cases where the LLS began with two structures, that with the highest molecular complexity score (discussed later) was selected as the origin. Routes leading to metal-based complexes were discarded as were those leading to targets of < 120 Da; these were frequently found to represent isotopically-labelled fragments or notable by-products from unsuccessful reactions.

Detailed information and analysis of the dataset is provided in the supplementary information, some aspects will be discussed here.

Route lengths (LLS) varied from 2 – 25 steps but became extremely sparse above approx. 14 steps. In total, the dataset comprises > 1.4m individual reactions, each of which was classified using NameRxn [1]. Automated assignment of reaction class was achieved with an overall success rate of 78%; on a per-target basis, however, 58% of synthetic routes had one or more unclassified transformation. Reactions from ACIE had the lowest classification rate (58%) and since novel reaction types and methodologies feature heavily in data from this source, this observation seems logical. We might expect reaction types from EJMC, JMC and OPRD to be broadly similar, so the lower degree of classification for reactions from OPRD (72%) is unexpected.

The distribution of the 10 main reaction classes from the dataset is shown in Figure 1 where variation by journal source was found to be minor. Aside from unclassified reactions, the predominance of alkylation/arylation and acylation transformations is clear. Protecting groups feature heavily with 43% of target structures and 30% of starting materials featuring a protecting group. Further evidence of the reliance on protecting groups is shown by the abundance of deprotection steps (Figure 1), the majority of which are cleavage of N-Boc groups. The discrepancy between deprotection and protection steps is presumably due to the prevalence of protected starting materials. A list of the top-20 most used specific reaction classes is provided in the supplementary information.

Figure 1. Distribution of reaction super-classes



All routes (2 – 25 steps LLS) included, n = 1.46m reactions

2. Fingerprint similarity as a measure of route progression

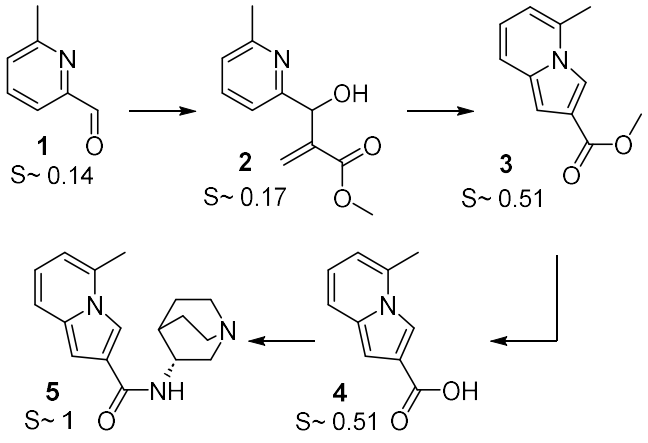
With a comprehensive dataset in place, we began our investigation with the use of molecular fingerprints. These are widely used in cheminformatics for applications including structure activity relationship (SAR) analysis, virtual library screening and computer aided synthesis planning (CASP). We were interested in studying their use as a measure of progress along a synthetic route -

between starting material and target - analogous to the use of reaction coordinates for single transformations.

Amongst the various fingerprint types, Morgan fingerprints [14] (comparable to the commonly-used enhanced connectivity fingerprint ECFP4 [15] when a bond radius of 2 is used) are popular for fast similarity comparison of small molecules and can be easily generated from a SMILES string. As a mathematical measure of similarity between two fingerprints, the Tanimoto coefficient is routinely used and yields values between 0 (no similarity) and 1 (identical) [16].

For a given synthetic route, the identity of the eventual target is known. A fingerprint of the target can therefore be generated and compared to the fingerprints of all preceding intermediates to the eventual starting material, giving a series of Tanimoto similarity values (S~). As an example selected randomly from the dataset, the synthesis of alpha-7 nAChR agonist **5** [17] is shown in Figure 2 with fingerprint similarity (S~) values versus the target **5**.

Figure 2. Synthesis of alpha-7 nAChR agonist



The starting material **1**, having only the 2-methylpyridine fragment in common with **5**, has the lowest S~ value and is the least "target-like" structure in the route. Transformation to **2** via Baylis-Hillman reaction gives a structure that is only slightly more similar to **5** (S~ 0.17). Although many of the C and O atoms present in **5** are introduced in this step, from a fingerprint perspective, few have the correct bond order since the ring system has not yet formed. The next step, cyclisation to give **3**,

displays a large change in target similarity (S~ 0.51) since the indolizine core is now present. Deprotection of the methyl ester to give **4**, has no impact on target similarity since **5** features neither a methyl ester nor carboxylic acid. The final amide formation introduces the rest of the target structure with a sizeable increase in similarity to give **5** (where necessarily S~ = 1).

Human inspection would quickly identify which of these transformations are productive and which are not. If we were considering ideality [8], for example, we would identify three construction steps and one protecting group manipulation, leading to an ideality of ¾ or 75% [8]. The same analysis could be achieved in an automated sense either by accurate assignment of reaction class or deduction from atom mapping along the route. Atom economy could also be calculated by considering atom mapping with the necessary reagents included[3]. The changes in fingerprint similarity across each step lead to largely the same conclusions: there are three steps where ΔS~ is positive and one step where it is zero. The magnitude of the ΔS~ values, particularly for the transformation of **1** to **2** highlight the subtleties of molecular fingerprints (or at least the Morgan fingerprint we are using here) in that the introduction of atoms present in the target structure has the added requirement of the correct bond order and cyclic/acyclic environment in order to have a positive effect.

This use of fingerprint similarity allows us to gauge progress along a given route but, on its own, is less useful for comparison of one route to another. As shown in xx, two unrelated transformations may have similar ΔS~ values by comparison to their respective route targets but represent markedly different degrees of structural change in comparison to each other.

We can supplement the changes in similarity with a measure of the magnitude of structural change taking place in a transformation. The changes in molecular weight (ΔMW) along the route might suffice here but, for our purposes of route assessment, we are interested in more than mass variation. Ideally, we are aiming to generate some measure of route efficiency related to cost, waste or time etc. Since this information is seldom available directly, we will use a molecular complexity

metric as a surrogate. There is an important assumption to recognise here in that we are assuming the "complexity" of a molecule is proportional to how easily it is obtained or synthesized, and therefore the implicit cost, time and waste. For the most part this seems reasonable in that "complex" molecules, where there is a variety of atom types, bond orders and ring systems are generally more challenging to obtain than "less complex" molecules. We must be mindful, however, that this assumption is insensitive towards complex molecules that are both naturally-occurring and readily available (eg, steroids, carbohydrates).

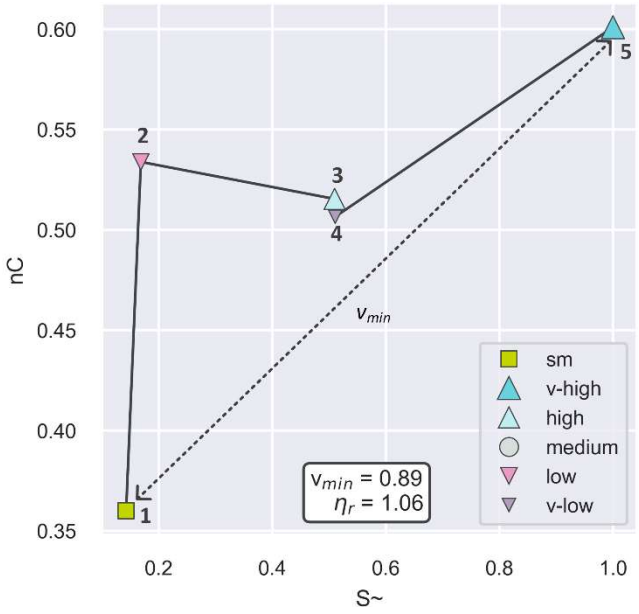
3. Similarity (S~) and complexity (C) as Cartesian coordinates

There are numerous molecular complexity metrics that might be used here and the choice will influence the behaviour of our analysis [18, 19, 20, 21, 22]. We will use a path-based complexity metric, C_{M^*} [23], that we have recently shown to be useful as a predictor of process mass intensity (PMI) [24]. A consideration of Böttcher [25] and spacial scores [26] as possible alternatives is provided in the supplementary information.

In the same way that we have used fingerprint similarity to gauge progress along a synthetic route, we will use this complexity measure to gauge the magnitude of structural change, *irrespective* of target similarity. As we will demonstrate shortly, these two measures can vary independently and need not be necessarily related.

In the dataset, the observed numerical ranges of S~ (0 – 1) and C_{M^*} (3.2 – 10.7) are different and in order to provide equal weight to similarity and complexity, C_{M^*} values were normalised and termed nC to give a range of 0 – 1. Thus, by plotting each structure as a pair of similarity/complexity coordinates, we can visualise the synthetic route from **1** to **5** (Figure 2) as shown in Figure 3.

Figure 3. Synthesis of alpha-7 nAChR agonist plotted as vectors



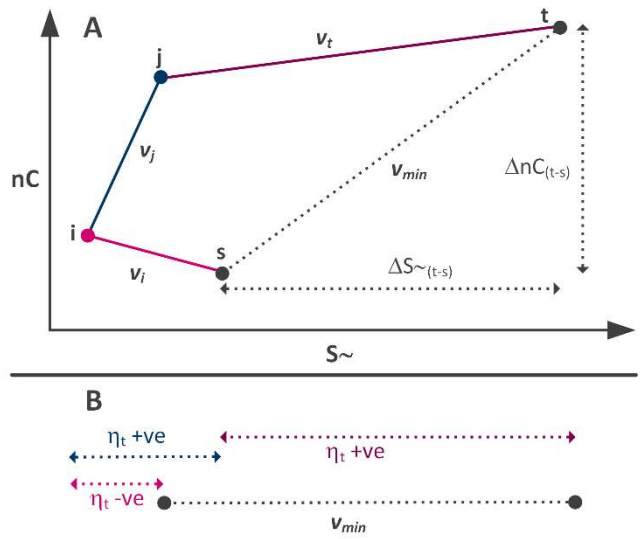
Each structure is plotted as a pair of similarity and complexity coordinates. Synthetic range (v_{min}) and range efficiency (η_r) are defined subsequently. Transformation efficiency is represented by marker hue, size and shape and also defined below.

When depicted as a series of head-to-tail vectors, the magnitude and direction of each vector gives us information about the productivity of each transformation in the context of the route. Since the starting material **1** is of lower complexity and (necessarily) of lower similarity than the target, the required direction of overall travel is up and left. The transformation of **1** to **2** adds significant complexity but, as we saw earlier, this complexity is not mirrored exactly in the eventual target. The indolizine formation (**2** to **3**) has already been highlighted as increasing target similarity significantly; there is a negative change in complexity associated since the number and types of atom environment and bond decreases due to the associated dehydration. The demethylation step (**3** to **4**) has minimal impact on both similarity and complexity and the final step sees a large increase in both.

4. Transformation efficiency

Also shown in Figure 3 are representations of the transformation efficiency (η_τ) of each step which is described in Figure 4 where a synthetic route (A) from starting material **s** to target **t** via intermediates **i** and **j**, is plotted as points on a similarity/complexity plane.

Figure 4. Derivation of scalar projections



On these axes, the overall synthetic range is represent by the vector v_{min} leading directly from **s** to **t**, which describes the required changes in similarity $\left(\Delta S\sim_{(t-s)}\right)$ and complexity $\left(\Delta nC_{(t-s)}\right)$ across the entire route. Theoretically, it also represents a "perfect" route from **s** to **t**, if such a transformation could be achieved in a single step. Each transformation is represented by a vector v_n describing the changes in similarity and complexity across the individual step. We can define the transformation efficiency η_t of each vector as the contribution it makes to the total change required for the route v_{min}. Since we are using simple Euclidean vectors, this is equal to the scalar projection [27] of v_n onto v_{min}; an example calculation is provided in the supplementary information.

Since the vector v_i (representing the transformation of **s** to **i**) has an opposing direction of travel to v_{mi}, the scalar projection, or transformation efficiency, is negative and effectively lengthens v_{min} (Figure 4 B) leaving increased synthetic "work" to be achieved in the remaining steps. Vectors v_j and

v_t are co-directional (but not parallel) with v_{min} leading to positive transformation efficiencies. Necessarily, the sum of the sum of the three scalar projections is equal to the synthetic range (magnitude of $\|v_{min}\|$) which allows us to assess the usefulness or efficiency of each step in the context of the overall route.

The representations of η_t shown in Figure 3 were calibrated using observed η_t values from the dataset by applying the categorical assessment shown in Table 1.

Table 1. Transformation efficiency (η_t) data

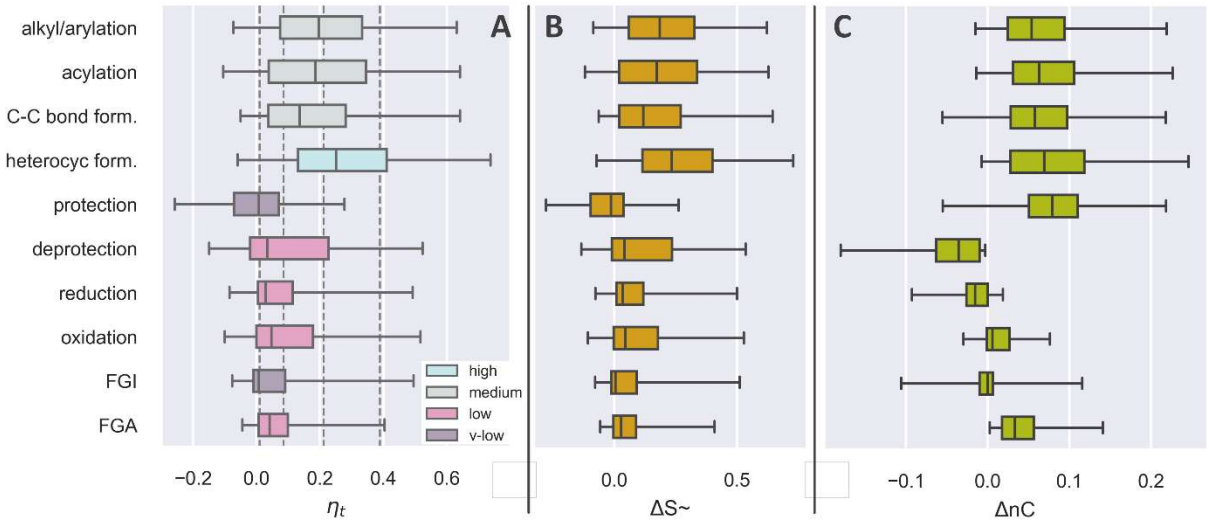
Measure	**Value**
Count	1.46m
Mean	0.198
Very low (< 20th percentile)	< 0.006
Low (< 40th percentile)	< 0.085
Medium (< 60th percentile)	< 0.211
High (< 80th percentile)	< 0.380
Very high (> 80th percentile)	> 0.380

Applying this analysis to the entire dataset and analysing by reaction class leads to some useful observations as shown in Figure 5. The trends are largely similar to those seen with the example above and in keeping with what we would expect as organic chemists. The ten reaction super-classes can be largely placed into three groups. Reactions from super-classes 1 – 4 (alkylation/arylation, acylation, C-C bond formation and heterocycle formation) are productive in the context of the routes in which they are utilised. They are almost exclusively associated with positive changes in similarity (Figure 5 B), complexity (Figure 5 C), and are classed as medium or, in the case of heterocycle formation, high median transformation efficiency (Figure 5 A). Super-classes 7 – 10 (reduction, oxidation, FGI and FGA) exhibit generally positive changes in similarity (since they

typically adjust existing atom environments to match the target structure), slightly positive or slightly negative changes in complexity and low or very low (FGI) median transformation efficiency.

The remaining two super-classes 5 and 6 (protection and deprotection) are distinct. Protection transformations are unique in effecting negative median change in similarity alongside the highest median change in complexity; overall, they are therefore classed as very low median transformation efficiency. Deprotections are the reverse with low (but skewed) median increase in target similarity, the greatest median decrease in complexity and overall low (and again skewed) median transformation efficiency.

Figure 5. Reactant to product changes in similarity (B), complexity (C) and transformation efficiency (A) by reaction super-class



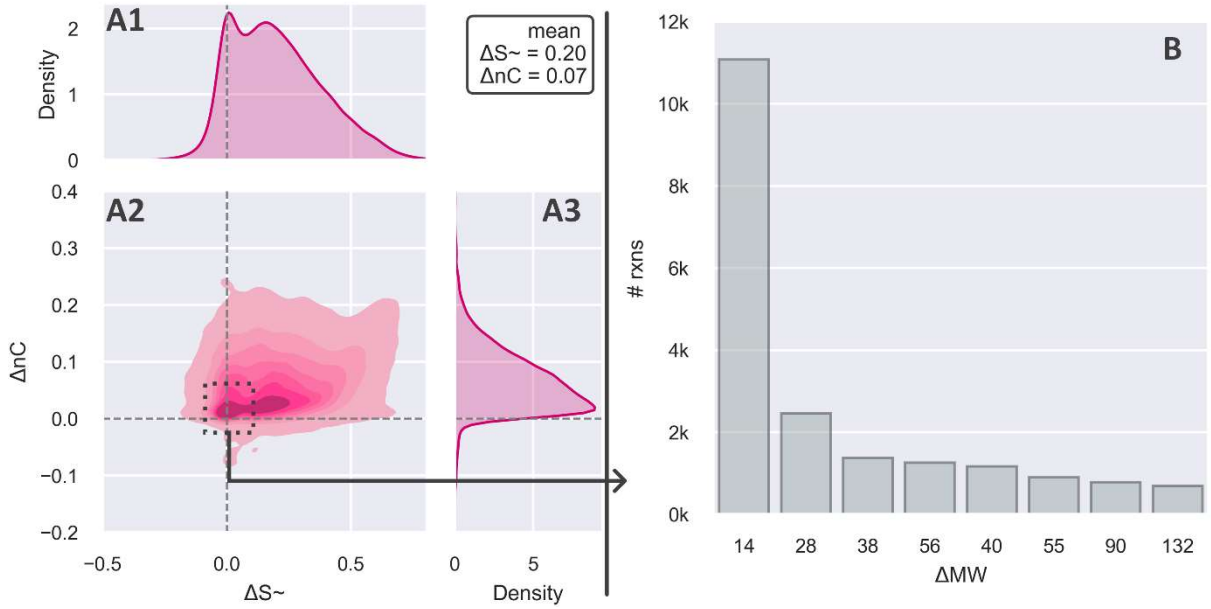
In order to minimise skew from very short or very long routes, data was compiled from routes of length 3 – 14 steps. Routes leading to targets featuring protecting groups were omitted, n = 533k reactions. Whiskers for individual reaction super-classes show 2nd and 98th percentiles, boxes show 25th, 50th and 75th percentiles. Transformation efficiency classes are shown by hue in chart C and the 20th, 40th, 60th and 80th percentiles of all reactions are marked as horizontal dashed lines.

This analysis seems logical for the most part but there are a number of associated subtleties and imperfections. As seen earlier, the use of a molecular fingerprint (or at least the Morgan fingerprint

we are using) to gauge similarity is sensitive towards bond order, hybridization, ring environment etc. If we imagine a two-step sequence where an amine RCH_2NH_2 is formed by i) addition of C≡N- to R-X and ii) reduction of R-C≡N, we would traditionally view the first step as productive and the second as non-productive. With the use of fingerprints, it is more likely that the first step will be seen as slightly productive and the second as more productive; calibration of η_t into five bands helps alleviate these irregularities. Furthermore, we should note that the fingerprint is susceptible to coincidental errors where, for example, addition of an -O^{t}Bu fragment by way of Boc protection might lead to a positive similarity change if the eventual target features a different (coincidental) -O^{t}Bu fragment.

Aside from these imperfections, there are notable benefits to the approach we are using in that "productivity" is not returned as a binary parameter where a transformation is either productive or non-productive. For example, in the case of heteroatom alkylation/arylation reactions (super-class 1), the transformation efficiency η_t is a continuous response that can be large (if a significant proportion of the target structure is being introduced) or small. The density plots in Figure 6 show the relationship between ΔS~ and ΔnC (A2) for super-class 1. Individual distributions for ΔS~ (A1) and ΔnC (A3) are also provided. The data is heavily clustered in the "productive" region (as seen earlier in Figure 5).

Figure 6. Similarity/complexity density plot (A) and change in MW (B) for heteroatom alkylation & arylation



Kernel density plot generated using the Seaborn library. Data was compiled from routes of length 3 – 14 steps, n = 223k reactions.

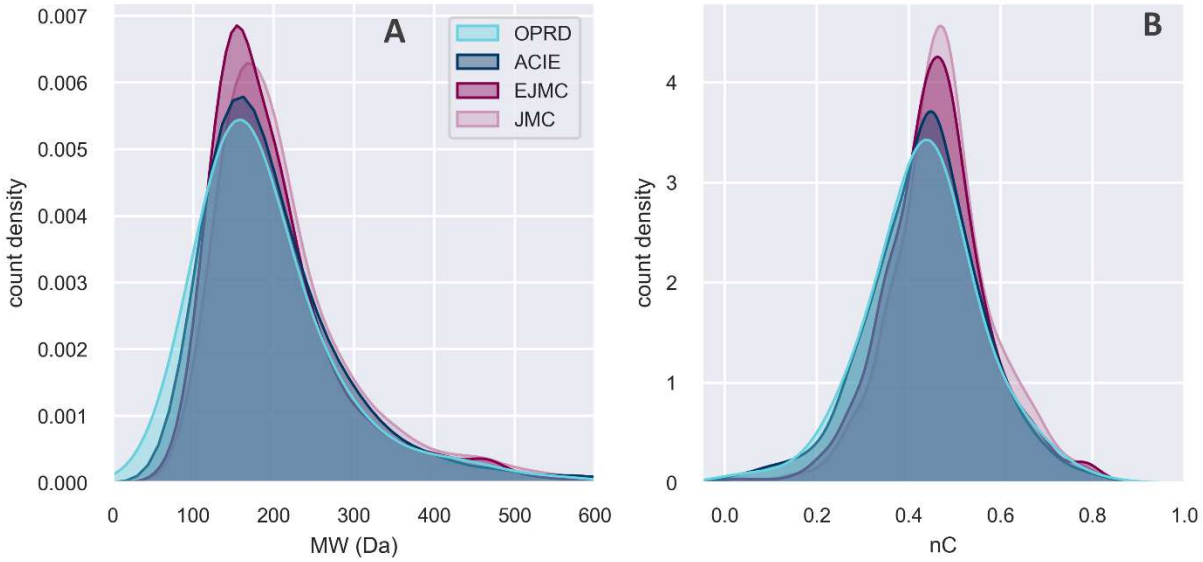
There is a concentration of low productivity reactions represented by the shoulder at ΔS~ = 0 (A1). Filtering the data to isolate this cluster, we can see the distribution of changes in MW for reactant to product (B). The vast majority of these reactions are associated with ΔMW +14 (which basic inspection identifies as methylation of alcohols, carboxylic acids etc) and +28 Da (ethylation). These transformations are deemed to be of low productivity as a result of either i) introduction of a fragment that is small in comparison to the rest of the molecule (ie, methylation or ethylation) or, ii) a structural change that is transient and not present in the final target structure meaning these reactions might more correctly be assigned as super-class 5 (protection).

Analogous plots of the other nine reaction super-classes are provided in the supplementary information.

5. Range efficiency

In the same way that transformation efficiency η_t was defined and calibrated against the dataset, we can assess the range efficiency η_r of a synthetic route. As discussed in the introduction, the comparison of synthetic routes (and particularly routes comprising different starting materials and/or targets) using step count alone is problematic due to a lack of convention concerning where to start counting steps. To illustrate this, the variability in starting material designation across the dataset is shown in Figure 7, where wide ranges are observed for MW (A) and complexity (B); these trends were observed across all four data sources.

Figure 7. Distribution of starting material molecular weight and complexity



Kernel density plot generated using the Seaborn library. Data was compiled from 356k routes of length 2 – 25 steps.

The definition of synthetic range v_{min} provided earlier (Figure 4) allows us to differentiate routes that are short because the starting material is complex from routes that are short due to high efficiency. In Figure 8, the synthetic range of routes in the dataset is shown to initially increase with step count then plateau. This leads to the somewhat pessimistic, but not wholly unexpected conclusion that routes longer than approximately 10 steps (LLS) do not, on average, incorporate a meaningful increase in synthetic range. Also shown, for each step count, are the 1st, 2nd and 3rd quartiles. The

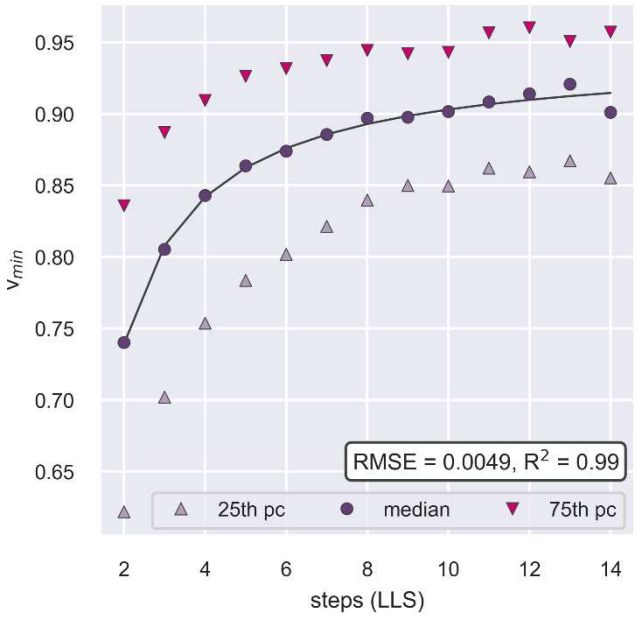
median value of v_{min} per step, termed v_{50}, was modelled, allowing us to define the route efficiency η_r as:

$$\eta_r = \frac{\|v_{min}\|}{v_{50}}$$

$$\text{where } v_{50} = 0.9435 - \frac{0.4033}{n} - \frac{0.0098}{n^2}$$

A value of $\eta_R = 1$ indicates a route has median range efficiency compared to routes of the same length from this dataset. Where $\eta_R > 1$, the range efficiency would be above average and where $\eta_R < 1$, below average. The calculated value for our example route to **5** (Figure 2), where a synthetic range of v_{min} = 0.89 is covered in 4 steps is η_r = 1.06; slightly above the median for routes of the same length.

Figure 8. Synthetic range (median, 25th/75th percentile) versus step count



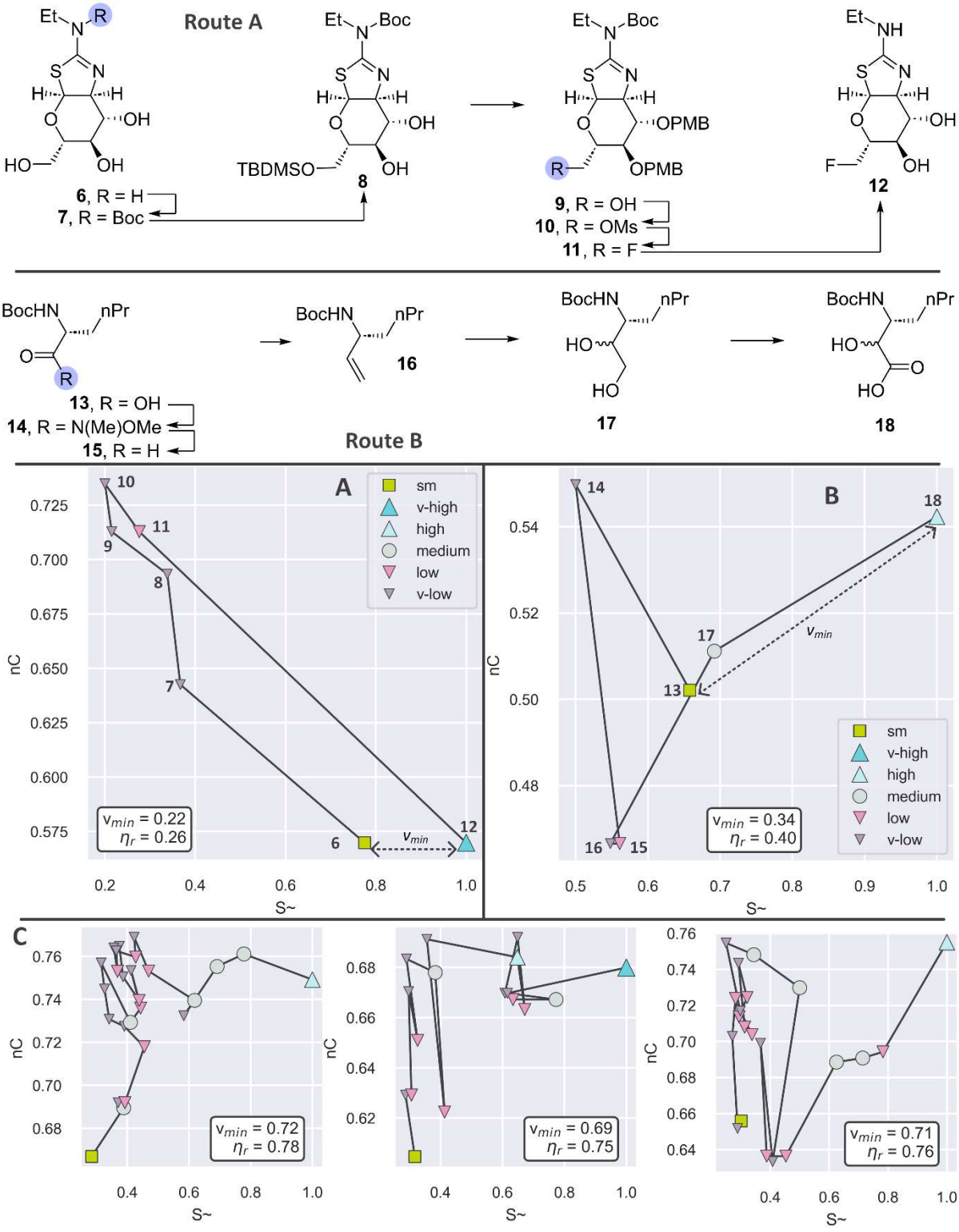
Data compiled from 249k routes of length 2 – 14 steps (LLS). Regression modelling carried out on median v_{min} values and inverse of step count (1/n) using scikit-learn[28].

Using this measure of range efficiency, we can quickly filter the dataset to show illustrative examples. The impact of protecting groups is exemplified in Figure 9 (route A) by the synthesis of O-GlcNAcase inhibitor **12** [29]. Small, poly-functionalised, chiral targets such as this one are notoriously challenging to synthesise and we should also stress that this is a divergent route to generate numerous analogues and is not optimised for this particular target. As shown on the vector plot (A), the synthetic range of the route is small, involving only exchange of hydroxyl to fluoride. In order to achieve this transformation regio-selectively, numerous orthogonal protecting groups were required (**6** through **9**), each associated with very-low η_t. The mesylation step (**9** to **10**) is an example of a very low efficiency acylation (super-class 2) and the ensuing fluorination (**10** to **11**) is of low efficiency since only a very small fragment of the target structure is introduced. Removal of the two protecting groups (**11** to **12**) is of very-high efficiency (the magnitude of this vector is three times greater than v_{min}). Again, this designation might seem counterintuitive but it is more than negated by the necessary protection steps; the overall range efficiency is low.

It is not only protecting groups that hamper range efficiency as shown in Figure 9 (route B). Homologation of amino acid **13**, another non-trivial synthetic task, delivered target **18** in 5 steps [30]. The vector diagram (B) shows the sizeable detours taken via the Weinreb amide **13**. The final two stages (**16** through **18**) are of higher efficiency but the synthetic range of the route is again low since overall, there is little change taking place in terms of molecular complexity; the range efficiency is similarly low.

For very long, challenging syntheses, the burden of high step-count becomes visually-striking, as shown by the (unattributed) routes in Figure 9 C. Even though the synthetic ranges are large, the observed vector path becomes highly erratic, with numerous transformations of low or very-low efficiency.

Figure 9.Examples of synthetic range and range efficiency

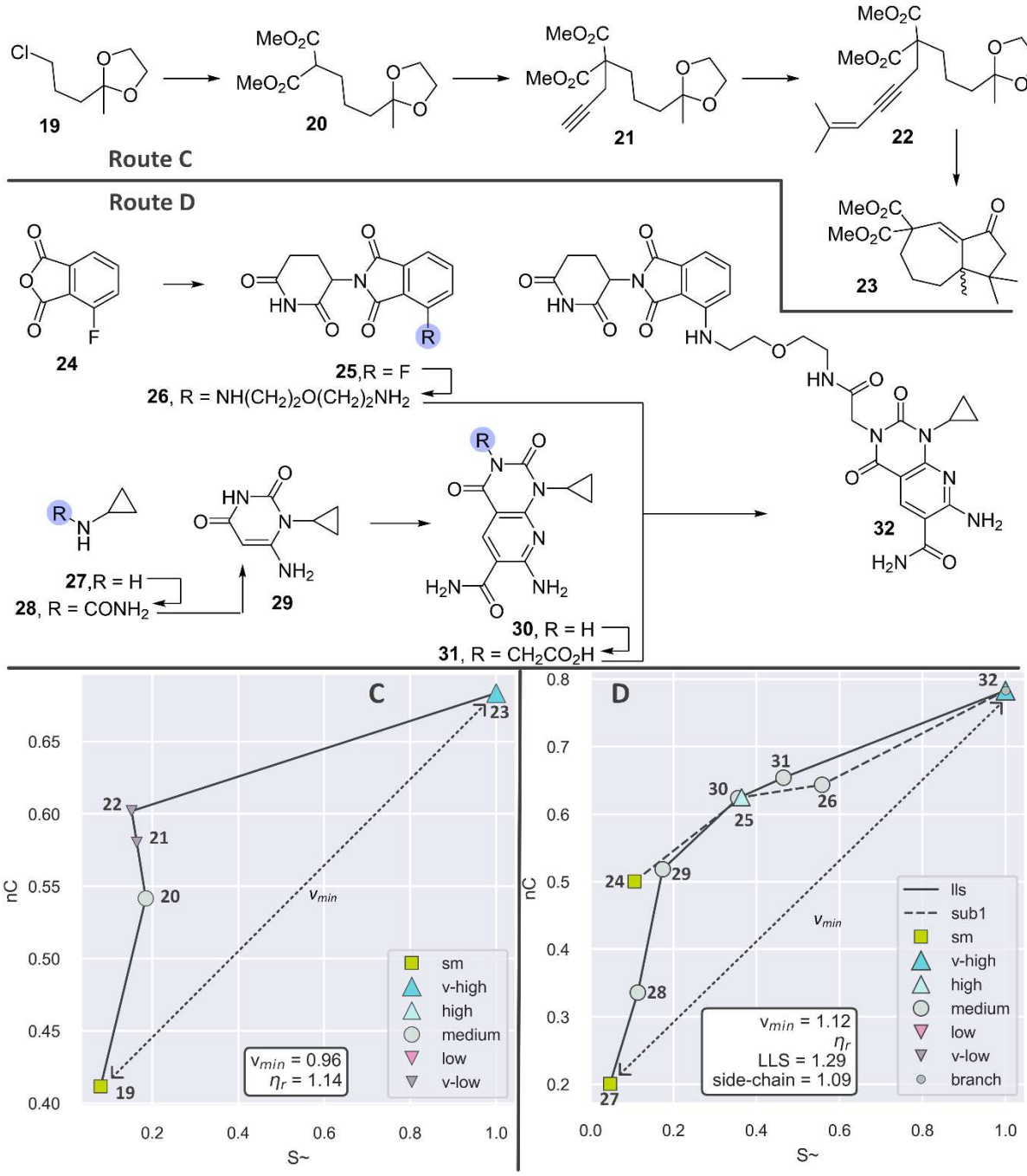


Routes C and D shown in Figure 10 demonstrate large synthetic range and high range efficiency. The synthesis of bicycle **23** (route C) again highlights the subtleties of the fingerprint analysis we are

using [31]. The vector plot (C) shows that whilst the first three steps (**19** through **22**) all add atoms present in the target structure, the first step (**19** to **20**) is of medium transformation efficiency and the subsequent two steps are of very low efficiency since the atom environments introduced do not exactly match those in the target. All of the "work" in terms of transformation efficiency is achieved in the final stage, where rings are formed and hybridization/bond orders reach the correct, final states. The range of the route (v_{min} = 0.96) is large for a four step route and the efficiency (η_r= 1.14) is above average. Basic inspection of the dataset suggests this kind of vector pattern is common for routes where complex fragments are constructed then completed via tandem or cascade transformations.

Route D (Figure 10) leading to eEF2K-targeting PROTAC molecule **33** covers a huge synthetic range (v_{min} = 1.12) in 5 steps (LLS) from amine **27** and the resulting range efficiency is very high (η_R = 1.29) [32]. The vector plot (D) shows every transformation is associated with medium, high or very high efficiency. This is also the first example we have shown that is convergent due to the presence of a parallel synthetic sequence.

Figure 10.Further examples of synthetic range and range efficiency

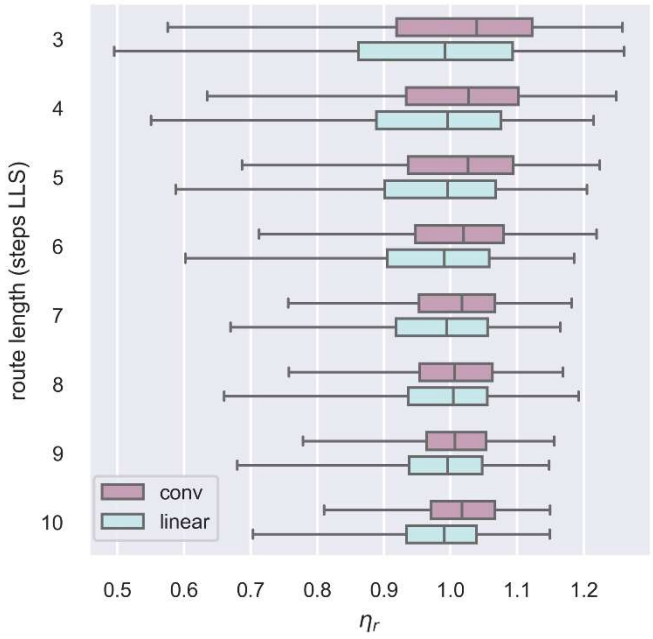


6. Linear & convergent routes

The designation of synthetic routes as linear or convergent is prone to the same inconsistencies as the designation of starting materials and the synthesis of fragments added along the LLS is frequently omitted if said fragment can be purchased or is deemed simple. The preceding example (Route D in Figure 10) was reported as a convergent route and was shown to be of high range

efficiency. This is not coincidental since it can be seen that convergent routes display higher median range efficiency than linear routes regardless of step count, as shown in Figure 11. Furthermore, inspection of the dataset suggests this is probably an underestimate as many of the linear routes of high range efficiency comprise complex incoming fragments and should more correctly be considered as convergent.

Figure 11. Range efficiency of linear and convergent syntheses



What has not been taken into account here is the efficiency of the synthetic side-chain leading up to the branch point; it is possible that a convergent route could feature an LLS with high range efficiency and a side-chain with low range efficiency. Exemplified in Figure 10 (D), we took a simple approach where the side-chain range efficiency is calculated separately between its starting material (**24**) and the branch point (**32**) with a step count of 3 and using the same calculation as for the LLS. Thus, in this case, we can see that both the LLS and side-chain are of above median range efficiency.

7. Final & future considerations

Since our methodology uses simple mathematical operations, there are various other analyses we might carry out. The addition of one (or more) extra axes with variable(s) to complement similarity and complexity would be straightforward, although vector plots of routes on three or more axes may be challenging. We have not discussed the impact of ordering in the sequence of similarity/complexity and transformation efficiencies along a synthetic route. We would instinctively suggest that similarity and complexity values should continually increase from starting material to target in order to minimise waste of valuable materials that inevitably occurs with yield losses along a sequence; we might similarly expect low-efficiency transformations to be better situated at the start of a route and high-efficiency transformations towards the end. All of these properties could be assessed using rank correlation metrics (eg, Pearson[33], Spearman[34]).

6. Conclusion

The use of vectors comprising similarity and complexity components to describe synthetic transformations and routes has been shown to be useful for visualising and quantifying a number of qualities that we look for as organic chemists. Using a large dataset, we have demonstrated that the transformation efficiencies (η_t) associated with particular reaction super-classes follow logical trends. The specific molecular fingerprint (Morgan) and complexity metric (C_{M^*}) we have used is somewhat nuanced in the way certain transformations are interpreted; we view this aspect as easily tuneable and it is likely that a different or custom-made fingerprint and/or complexity metric could be used, following the same principles. We have also described how the synthetic range of a route (v_{min}) can be represented as a vector to alleviate inconsistencies with step-counting and starting material designation. The range efficiency (η_r) describes how effectively a given route (of length 2 – 14 steps LLS) traverses the required changes in similarity and complexity as a function of step count and we have shown how synthetic routes comprising different starting materials and targets can be rapidly assessed, filtered and compared.

The definitions of transformation and range efficiency provided are calibrated against a dataset of > 350k literature syntheses, the majority of which are pharmaceutical in origin. This calibration might not be appropriate for other synthetic applications where significantly shorter (eg, fine chemicals) or longer (eg, natural products) routes are employed, although we would expect the same principles to be relevant.

It should be stressed that the methodology described here, derived only from chemical structures and route topography is wholly theoretical and will always be inferior to real, empirical data such as cost, time and waste. Obtaining reliable empirical data for known transformations is problematic however and, in the case of unknown or theoretical transformations, Hendrickson's observation made in 1976 ("when planning an organic synthesis it is presently impossible to predict the yields of individual reactions, or indeed even whether they will succeed or fail") remains pertinent today. Thus, we believe the methodology described here, which is highly amenable to machine interpretation, will be useful wherever automated assessment of synthetic transformations and routes is applied.

7. Supplementary information

Further statistical analyses of the dataset are provided along with discussion of notable aspects relating to targets, routes and reaction super/sub-classes. Equations for the derivation of vector magnitudes, dot products and scalar projections are described in further detail. Due to licensing restrictions, we are unable to share the dataset used in this manuscript. As an alternative, we have provided a smaller dataset of patent syntheses along with a Jupyter notebook and supporting code to demonstrate how synthetic routes can be compiled and analysed.

8. Conflict of interests

For the duration of this study, GPH and SG were both employees of and shareholders in AstraZeneca, who funded the research.

9. References

1 https://www.nextmovesoftware.com/namerxn.html. Accessed December 2023
2 Kraut H, Eiblmaier J, Grethe G, et al (2013) Algorithm for reaction classification. Journal of Chemical Information and Modeling 53:2884–2895. doi: 10.1021/ci400442f

3 Trost B (1991) The atom economy—a search for synthetic efficiency. Science 254:1471–1477. doi: 10.1126/science.1962206
4 Wender PA, Croatt MP, Witulski B (2006) New reactions and step economy: The total synthesis of (±)-salsolene oxide based on the type II transition metal-catalyzed intramolecular [4+4] cycloaddition. Tetrahedron 62:7505–7511. doi: 10.1016/j.tet.2006.02.085
5 Wender PA, Verma VA, Paxton TJ, Pillow TH (2007) Function-oriented synthesis, step economy, and Drug Design. Accounts of Chemical Research 41:40–49. doi: 10.1021/ar700155p
6 Wender PA, Miller BL (2009) Synthesis at the molecular frontier. Nature 460:197–201. doi: 10.1038/460197a
7 Burns NZ, Baran PS, Hoffmann RW (2009) Redox economy in organic synthesis. Angewandte Chemie International Edition 48:2854–2867. doi: 10.1002/anie.200806086
8 Gaich T, Baran PS (2010) Aiming for the ideal synthesis. The Journal of Organic Chemistry 75:4657–4673. doi: 10.1021/jo1006812
9 Hendrickson JB (1977) Systematic synthesis design. 6. yield analysis and convergency. Journal of the American Chemical Society 99:5439–5450. doi: 10.1021/ja00458a035
10 Lin A, Dyubankova N, Madzhidov TI, et al (2021) Atom‐to‐Atom mapping: A benchmarking study of popular mapping algorithms and consensus strategies. Molecular Informatics. doi: 10.1002/minf.202100138
11 https://www.reaxys.com. Accessed December 2023
12 Mo Y, Guan Y, Verma P, et al (2021) Evaluating and clustering retrosynthesis pathways with learned strategy. Chemical Science 12:1469–1478. doi: 10.1039/d0sc05078d
13 Genheden S, Bjerrum E (2022) Paroutes: Towards a framework for benchmarking retrosynthesis route predictions. Digital Discovery 1:527–539. doi: 10.1039/d2dd00015f
14 Morgan HL (1965) The generation of a unique machine description for chemical structures-a technique developed at Chemical Abstracts Service. Journal of Chemical Documentation 5:107–113. doi: 10.1021/c160017a018
15 Rogers D, Hahn M (2010) Extended-connectivity fingerprints. Journal of Chemical Information and Modeling 50:742–754. doi: 10.1021/ci100050t
16 Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics. doi: 10.1186/s13321-015-0069-3
17 Xue Y, Tang J, Ma X, et al (2016) Synthesis and biological activities of indolizine derivatives as alpha-7 nachr agonists. European Journal of Medicinal Chemistry 115:94–108. doi: 10.1016/j.ejmech.2016.03.016
18 Bertz SH (1981) The first general index of molecular complexity. Journal of the American Chemical Society 103:3599–3601. doi: 10.1021/ja00402a071
19 Hendrickson JB, Huang P, Toczko AG (1987) Molecular complexity: A simplified formula adapted to individual atoms. Journal of Chemical Information and Computer Sciences 27:63–67. doi: 10.1021/ci00054a004
20 Li J, Eastgate MD (2015) Current complexity: A tool for assessing the complexity of organic molecules. Organic & Biomolecular Chemistry 13:7164–7176. doi: 10.1039/c5ob00709g
21 Sheridan RP, Zorn N, Sherer EC, et al (2014) Modeling a crowdsourced definition of molecular complexity. Journal of Chemical Information and Modeling 54:1604–1616. doi: 10.1021/ci5001778
22 Coley CW, Rogers L, Green WH, Jensen KF (2018) SCScore: Synthetic complexity learned from a reaction corpus. Journal of Chemical Information and Modeling 58:252–261. doi: 10.1021/acs.jcim.7b00622
23 Proudfoot JR (2017) A path based approach to assessing molecular complexity. Bioorganic & Medicinal Chemistry Letters 27:2014–2017. doi: 10.1016/j.bmcl.2017.03.008
24 Angelini L, Coomber CE, Howell GP, et al (2023) Cumulative complexity meta-metrics as an efficiency measure and predictor of process mass intensity (PMI) during synthetic route design. Green Chemistry 25:5543–5556. doi: 10.1039/d3gc00878a
25 Böttcher T (2016) An additive definition of molecular complexity. Journal of Chemical Information and Modeling 56:462–470. doi: 10.1021/acs.jcim.5b00723
26 Krzyzanowski A, Pahl A, Grigalunas M, Waldmann H (2023) Spacial score—a comprehensive topological indicator for small-molecule complexity. Journal of Medicinal Chemistry 66:12739‐12750. doi: 10.1021/acs.jmedchem.3c00689
27 https://en.wikipedia.org/wiki/Scalar_projection. Accessed February 2024.
28 https://scikit-learn.org/stable/. Accessed February 2024.

29 Selnick HG, Hess JF, Tang C, et al (2019) Discovery of MK-8719, a potent O-glcnacase inhibitor as a potential treatment for Tauopathies. Journal of Medicinal Chemistry 62:10062–10097. doi: 10.1021/acs.jmedchem.9b01090
30 Velmourougane G, Harbut MB, Dalal S, et al (2011) Synthesis of new (−)-bestatin-based inhibitor libraries reveals a novel binding mode in the S1 pocket of the essential malaria M1 metalloaminopeptidase. Journal of Medicinal Chemistry 54:1655–1666. doi: 10.1021/jm101227t
31 Escalante L, González‐Rodríguez C, Varela JA, Saá C (2012) Tandem Brønsted acid promoted and Nazarov carbocyclizations of enyne acetals to hydroazulenones. Angewandte Chemie International Edition 51:12316‐12320. doi: 10.1002/anie.201205823
32 Liu Y, Zhen Y, Wang G, et al (2020) Designing an eef2k-targeting PROTAC small molecule that induces apoptosis in MDA-MB-231 cells. European Journal of Medicinal Chemistry 204:112505. doi: 10.1016/j.ejmech.2020.112505
33 https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. Accessed January 2024.
34 https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient. Accessed January 2024.