A database tool for process chemists and chemical engineers to gauge the material and synthetic efficiencies of synthesis plans to industrially important targets*

John Andraos‡

CareerChem, 504-1129 Don Mills Road, Don Mills, ON M3B 2W4, Canada

Abstract: After 20 years of green chemistry research, a complete algorithm for the determination of material and synthetic strategy efficiencies for synthesis plans to any chemical target has been achieved. This paper presents the first announcement of a comprehensive database consisting of green metrics calculations for 1060 plans to 220 targets of interest to the chemical industry in the following categories: commodity industrial chemicals, pharmaceuticals, agrichemicals, dyestuffs and colorants, natural products, flavorings, fragrances, and sweeteners, and molecules of theoretical interest. Data mining of the original literature covered the period 1828–2010. A summary of trends in achieving green chemistry strategies is presented, including an unbiased method of ranking plans using a suite of parameters, ring construction strategies, and implications on the development of new kinds of smart structure search databases. The take-home message is that targeted optimization is a multivariable problem that requires synergistic maximization and minimization of key variables. Problems in the reporting of chemical syntheses in scientific journals and patents are discussed as well as setting guidelines for their standardization and normalization. The merits of spreadsheet tools are presented from decision making in route selection all the way to fast and accurate proofreading of the final plan chosen.

Keywords: agrichemicals; atom economy; dyestuffs; E-factor; flavorings; fragrances; green metrics; optimization; pharmaceuticals; synthesis strategy; sweeteners.

INTRODUCTION

The global amount of waste produced from a chemical process is easily determined by taking the difference between the sum of masses of all input materials and the mass of desired product. However, in order to minimize this global waste with the aim of devising an optimum synthesis plan for a given target molecule, one needs to first identify and itemize its constituents. For any individual chemical reaction, these include unreacted and/or excess reagents, byproducts produced as a consequence of making the desired target product, side products produced as a consequence of competing side reactions occurring in parallel to the intended reaction, reaction solvents, and all auxiliary materials used in the work-up and purification phases in the course of carrying out the said reaction. Knowledge of what constitutes waste for a chemical process requires the full understanding of the chemistry involved in every reaction.

*Paper based on a presentation made at the 3rd International Conference on Green Chemistry (ICGC-3), Ottawa, Canada, 15–18 August 2010. Other presentations are published in this issue, pp. 1343–1406.
‡E-mail: c1000@careerchem.com
reaction step. This is necessary before attempting to minimize it. In order to carry out this task, one needs the following for each reaction in a synthesis plan: a fully balanced chemical equation, knowledge of the reaction mechanism, knowledge of reaction kinetics and thermodynamics, and knowledge of all chemical structures involved from reactants to products. The exercise of optimizing a chemical process is in fact doing green chemistry. In this vein, the well-known fundamental metrics developed to gauge individual reaction performance are atom economy (AE) [1] and E-factor [2]. The definitions of these metrics are based on positive and negative perspectives with respect to producing the desired target product, respectively. The mathematical connection between them is, of course, based on the balanced chemical equation and the law of conservation of mass discovered by Lavoisier [3]. This insight leads directly to the fundamental relationship in eq. 1 for reaction mass efficiency (RME) of any chemical reaction [4].

\[
RME = \frac{\varepsilon (AE)(1/SF)(MRP)}{1/(1 + E)} = \frac{1}{1 + E}
\]  

where \(\varepsilon\) is the reaction yield with respect to limiting reagent, AE is atom economy, \(1/SF\) is the inverse of the stoichiometric factor which designates the consumption of excess reagents (\(SF = 1\) means that no excess reagents are used, \(SF > 1\) means that at least one reagent is used in excess), MRP is the material recovery parameter, which designates the consumption of all auxiliary materials other than reactants, and \(E\) is the E-factor, which is the ratio of mass of overall waste to mass of desired product.

Over the last decade, green metrics has become a recognized area of vigorous research and has been extensively reviewed [5]. Figure 1 shows a new paradigm for the determination of material and synthetic efficiencies for any synthesis plan. A key feature of this new paradigm is the visual representation of efficiency parameters so that strengths and weaknesses of a synthesis strategy may be easily identified. From a set of balanced chemical equations and their associated detailed experimental procedures in which all masses of materials used are disclosed, a reaction performance spreadsheet algorithm is used to construct a set of radial pentagons [6] for all chemical reactions in a synthesis plan. These pentagons are visual representations of the five parameters given in eq. 1. Also, an accompanying synthesis tree [7] is constructed which shows the number, kind, and molecular weights of input materials, the number of branches, the number of points of convergence if any, molecular weights of all intermediates, and reaction yields of all reaction steps in a simplified diagram. The results of the radial pentagons and the synthesis tree diagram are then fed into a synthesis plan spreadsheet algorithm [8] which determines the overall material and synthesis efficiencies of the plan. Various distribution profiles are constructed for a variety of parameters, a target bond map is constructed that shows which bonds are made in the desired product and in which steps, and a radial hexagon [8] is constructed that describes the overall plan performance based on overall AE, overall reaction yield (with respect to the longest branch), overall kernel RME (based on reaction yields and byproduct wastes only), fraction of non-sacrificial reagents used, fraction of kernel waste arising from target bond-forming reactions, and degree of convergence. Spreadsheet algorithms and synthesis tree analysis impart the following advantages, as they

- provide a proofreading tool for patent examiners of patent applications and reviewers of manuscripts to screen for errors in experimental procedures;
- provide an intelligent in-depth critique of plans;
- allow unbiased comparisons between plans;
- provide concrete proof that a newly reported plan is different and better by material consumption and strategic design than prior reported plans;
- provide proof of novelty in improvements to process patents; and
- provide a powerful tool for scientists to do “what if” analyses and see the instant results of their choices during the design phase of a synthesis before going to the lab.

For any synthesis plan, the total number of steps may be subdivided into two categories: those that produce target bonds (target bond-forming reactions) and those that do not (sacrificial reactions).
From a strategy perspective, the objective is to simultaneously minimize the overall number of steps and to maximize the number of target bond-forming reactions. The same double optimization holds true from a waste perspective. The task is to simultaneously minimize global waste and to minimize the waste contribution from sacrificial reactions. If done successfully, the result is an overall "green" synthesis from a material consumption perspective.

**DATABASE HIGHLIGHTS AND TRENDS**

Table 1 summarizes the size of the synthesis database [9] constructed for the elucidation of "green" synthesis strategies. From an education perspective, the database serves several purposes: it provides a historical record of synthesis plans to important target molecules; it is an excellent resource for teachers and students to learn about organic synthesis; and it is a superb resource for green chemistry. From a research perspective, the database is a powerful tool in answering the following important questions: (a) Has a given target compound been made before? (b) Which synthesis plan is currently optimal to make it and why? (c) Is a given proposed plan competitive with respect to material efficiency with what has been done before? (d) Is a given proposed plan novel in terms of synthesis strategy compared to what was done before?

**Table 1** Summary of synthesis database target molecules and plans.

<table>
<thead>
<tr>
<th>Target type</th>
<th>Number of targets</th>
<th>Number of plans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural products</td>
<td>82</td>
<td>528</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>43</td>
<td>243</td>
</tr>
<tr>
<td>Agrichemicals</td>
<td>25</td>
<td>59</td>
</tr>
<tr>
<td>Dyestuffs and colorants</td>
<td>21</td>
<td>62</td>
</tr>
<tr>
<td>Flavorings, fragrances, and sweeteners</td>
<td>12</td>
<td>73</td>
</tr>
<tr>
<td>Molecules of theoretical interest</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>Miscellaneous industrial commodity chemicals</td>
<td>25</td>
<td>58</td>
</tr>
<tr>
<td>Totals</td>
<td>221</td>
<td>1065</td>
</tr>
</tbody>
</table>

© 2011, IUPAC  
Each synthesis plan in the database is evaluated on the basis of the paradigm and spreadsheet algorithms described above. The results of each plan are summarized in an individual electronic file consisting of the following parts: (a) compound name; (b) plan literature references including those for starting materials; (c) synthesis plan scheme showing balanced equations with byproducts and reaction yields for each step; (d) synthesis tree diagram; (e) list of key performance metrics including plan type (linear or convergent), number of branches, number of reaction stages, number of reaction steps, number of input materials, overall AE, overall kernel RME, overall yield based on longest branch, kernel mass of waste to produce 1 mol of target molecule, total mass of waste to produce 1 mol of target molecule, E-kernel, E-excess, E-auxiliaries, E-total, molecular weight first moment building up parameter, molecular weight fraction of sacrificial reagents used, hypsicity (oxidation level) index, degree of convergence, and degree of asymmetry; (f) target bond map and list of reagents that end up in target molecule structure; (g) list of graphical profiles including radial hexagon, molecular weight building up profile, AE distribution, % kernel waste distribution, % total waste distribution, reaction yield distribution, kernel RME distribution, hypsicity profile, and target bond-forming reaction profile; and (g) radial pentagons for each reaction. For target products with multiple plans, summary tables of plan performances are ranked by kernel waste production due to byproducts and unreacted starting materials, total waste production due to all sources, and E-factor breakdowns. In addition, from the set of target bond maps a Tanimoto similarity analysis [10] is performed to elucidate which pairs of plans are similar and dissimilar with respect to synthesis strategy. A Tanimoto index close to 1 indicates high similarity, whereas a value close to 0 indicates high dissimilarity. For target molecules that involve rings, their associated plans were also evaluated by graphical means to enumerate their ring construction strategies. Patterns of ring construction were identified and grouped by type. The Tanimoto and the ring construction strategy analyses suggest new opportunities to construct smart databases for structure searches based on specified target synthesis bonds. Such databases would help researchers answer the question of novelty in planning a synthesis based on a retro-synthetic analysis when comparing a proposed methodology with prior published work to the same target molecule.

Some basic statistics for the database are summarized below:

- Over 12,000 reactions were evaluated.
- 65% of the plans were linear, and 35% were convergent with at least two branches.
- 54% of molecular targets were of interest to academics, and 46% were of interest to industrial chemists.
- 24% of plans were performed in industrial labs, and 76% in academic labs.
- The oldest plan analyzed is the Graebe–Liebermann 1869 plan for alizarin [11].
- The newest plan analyzed is the Aggarwal 2010 plan for quinine [12].
- The shortest plan analyzed is the one-step Sandborn 1941 plan to (–)-menthone from (–)-menthol [13].
- The longest plan analyzed is the 85-step Ley 2008 plan for azadirachtin [14,15].
- The overall best performing plan is the Givaudan–Roure 1998 plan for vanillin from ferulic acid [16,17].
- The worst performing plan is the Willstätter G1 1901 plan for tropinone [18,19].

Figure 2 shows pie charts depicting various categories of plan material and synthesis efficiency performance for academic and industrial plans based on the molecular weight fraction of sacrificial reagents used [f(sac)] and the fraction of reaction steps that form target bonds [f(tbd)]. It can be seen from these diagrams that the number of truly green plans is very low. Gratifyingly, the number of very poor performing plans is also low. The positive aspect is that the vast majority of synthesis plans in both categories have a strong potential of becoming “green” once their shortcomings are identified. Up to now, chemists’ efforts in synthesizing compounds have largely been to prove chemical structures and to test the feasibility of various synthetic methodologies. This is especially true of academic plans that place greater emphasis on novelty and proof of structure. Though industrial chemists have a long track record...
in optimizing reaction and plan performance as substantiated by the data shown in the second pie chart, the concept of directed optimization in a green chemistry context is a recent development for all types of synthetic chemists. The results shown in the pie charts suggest that much work remains to be done.

**RADIAL PENTAGON EXAMPLE**

An illustration of the radial pentagon analysis is shown in Fig. 3 for the three-component coupling reaction of an aldehyde, alkyne, and amine [20]. The procedure in *Organic Syntheses* disclosed masses of all input materials used in all categories, which is a rare occurrence in the chemistry literature, but a necessary one for the true and rigorous evaluation of “greenness” of a given reaction. The amounts of reagents used were: cyclohexanecarboxaldehyde (337 mg, 3.009 mmol), ethynyl trimethylsilane (295 mg, 3.01 mmol), and dibenzylamine (592 mg, 3.005 mmol). From these data, the dibenzylamine is the limiting reagent. The amounts of catalytic and auxiliary materials used were: cuprous bromide (21.5 mg), (S)-quinap (72.5 mg), and molecular 4 A sieves (1.5 g). The reaction solvent was 9 mL (7.79 g) of toluene. The work-up solvents used were 50.15 mL (35.5 g) diethyl ether and 14.85 mL (9.3 g) n-pentane. For the purification, column chromatography using 50 g silica gel and 500 mL of a 99:1 v/v n-pentane:diethyl ether solvent mixture (313.41 g) was used as eluent. The radial pentagon spreadsheet algorithm yields the following results: AE = 95.6 %, reaction yield = 91.5 %, SF = 1.0008, MRP = 0.00292, and RME = 0.225 %. The reaction has a high AE since it produces water as the only byproduct that has a very low molecular weight compared to the target product. The E-factor breakdown given in Fig. 3 shows that the auxiliary material contribution, E-aux, is by far the greatest component of E-total. The contribution from excess reagents is small since SF is only slightly larger than 1. The E-kernel component from byproducts is the second highest contributor to E-total. The radial pentagon shows that if auxiliary materials are recovered the overall RME dramatically increases to 87.4 %.

© 2011, IUPAC
An illustration of the radial hexagon concept is shown in Fig. 4 for the Takasago G1 synthesis plan for (–)-menthol from myrcene [21] depicted in Scheme 1. The overall AE is 61.4 %, the overall yield over the 5 steps is 32.2 %, and the overall kernel RME is 20.2 %. The reactants used that in whole or in part end up in the (–)-menthol structure are myrcene, hydrogen (H₂), and 2 molecules of water. All other reagents appearing in Scheme 1 are sacrificial, which amounts to 31.5 % of all reagents used on a molecular weight basis. All of the kernel waste originates from target bond-forming reactions and all steps produce target bonds. The E-factor breakdown is as follows: E-kernel = 3.94, E-excess = 16.20, and E-aux = 290.40. From Fig. 4, an overall metric that conveniently encompasses all six parameters is the vector magnitude ratio (VMR) of all components found for the plan relative to those for a perfectly green plan with all parameters equal to one as shown in eq. 2.

$$VMR = \left(\frac{1}{\sqrt{6}}\right) \left[\sum_{j=1}^{6} r_j^2\right]$$

(2)

This parameter may be used to rank plans according to material and synthetic strategy efficiencies. By this metric, the Takasago G1 plan ranks 5th out of 7 plans [9].

Fig. 3 Radial pentagon for three-component coupling reaction shown involving an aldehyde, an alkyne, and an amine [20]. Note that dashed line is hidden behind “partial reclaiming” line.
The synthesis of (+)-sertraline hydrochloride, an antidepressant manufactured by Pfizer, is a good illustrative example of synthesis plan analysis found in the database. The third-generation plan [22] shown in Fig. 5 was awarded a U.S. Presidential Green Chemistry Award by the Environmental Protection Agency.

**SYNTHESIS PLAN ANALYSIS EXAMPLE**

The synthesis of (+)-sertraline hydrochloride, an antidepressant manufactured by Pfizer, is a good illustrative example of synthesis plan analysis found in the database. The third-generation plan [22] shown in Fig. 5 was awarded a U.S. Presidential Green Chemistry Award by the Environmental Protection Agency.

**Scheme 1** Takasago G1 plan for (–)-menthol [21]. Reaction yields for each step and molecular weights of intermediates appear above and below structures, respectively.
Agency in 2002 [23]. Figure 6 shows the corresponding synthesis tree diagram. It is straightforward to see that the plan is linear with 5 reaction steps and involves 14 input materials. As an illustrative example of the versatility of this kind of diagram, the determination of the required stoichiometric amounts of (R)-(−)-mandelic acid and 1,2-dichlorobenzene may be determined at once by multiplying the molecular weight of each reagent by the required target mole scale of (+)-sertraline hydrochloride product and dividing this result by the product of reaction yields that connect the product node and each of the respective reactant nodes as indicated in the diagram. Hence, to produce 1 mole of (+)-sertraline hydrochloride, Fig. 6A shows that 1/(0.40*0.81) = 3.09 mol or 152*3.09 = 469.1 g of (R)-(−)-mandelic acid are required. Similarly, Fig. 6B shows that 1/(0.92*0.92*0.91*0.40*0.81) = 4.01 mol or 146.9*4.01 = 588.7 g of 1,2-dichlorobenzene are required. Figure 7 shows the corresponding radial pentagons for all reactions in the plan. From these pentagons, the following bottlenecks in the plan may be easily identified: steps 1, 2, and 5 have the lowest AE; step 4 has the lowest reaction yield due to the automatic 50% loss of material in the classical resolution; steps 1, 2, and 3 use a significant amount of excess reagents; and steps 3, 4, and 5 use a significant amount of auxiliary solvent materials. Even if the reactions were not given, these bottlenecks may be readily identified by examining the shapes of the pentagons since poor performance in any one of the parameters is associated with a distortion of the pentagon toward the center. Good performance, on the other hand, is indicated by values of parameters that lie at or near the perimeter. The radial hexagons in Fig. 8 show the Pfizer G3 plan performance in context with other published plans [24]. Based on the VMR as defined by eq. 2, this plan ranks second among 9 plans. Figures 9 and 10 show the percent kernel and percent total waste distribution profiles.
Fig. 6 Synthesis tree diagram for Pfizer G3 plan for (+)-sertraline hydrochloride: (A) chain of reaction yields pertaining to amount of (R)-(−)-mandelic acid required, (B) chain of reaction yields pertaining to amount of 1,2-dichlorobenzene required.

Fig. 7 Radial pentagons for various synthesis plans for (+)-sertraline hydrochloride: step 1 (Friedel–Crafts acylation), step 2 (reduction–lactonization), step 3 (Friedel–Crafts acylation–cyclization), step 4 (imination–hydrogenation–classical resolution sequence), and step 5 (free base–hydrochloride salt formation sequence).
respectively. For the majority of cases, a pairwise comparison of these bar graphs for any given plan shows that the ranking order of reaction steps producing waste changes when auxiliary materials are included in the computation. Figure 11 shows the ranking of plans according to E-factor and fraction of waste originating from target bond-forming and non-target bond-forming reactions. A steady progression of overall waste reduction is observed from the first- to the third-generation Pfizer plans. Also, the proportion of the kernel waste arising from sacrificial reactions vanishes in the third-generation plan. These results are consistent with the green chemistry award bestowed on Pfizer, thus strengthening its credibility. The E-factors ranked chronologically for the plans show a steady and significant reduction over the years except for recently disclosed plans [24f,h], though most of the waste in that plan is aqueous in origin. In all plans the E-aux contribution is highest, which is typical of almost all plans examined in the database. The target bond maps are shown in Fig. 12. From these diagrams, it may be deduced that four categories of ring construction strategies have been employed to date in the synthesis of this molecule as summarized in Fig. 13. The Gedeon–Richter and Zhao plans begin with 1-naphthol which already has rings A and B established. The Buchwald, Pfizer G2, and Pfizer G3 plans make use of a [4 + 2] cycloaddition via Friedel–Crafts chemistry in constructing ring B. The Lautens plan involves a standard Diels–Alder type [4 + 2] cycloaddition that results in a bicyclic intermediate, which is then reductively opened. The Pfizer G1 plan employs a [6 + 0] cycloaddition again using Friedel–Crafts chemistry. All plans begin with starting materials having rings A and C preformed, and hence the objective in all plans is the construction of ring B with both 1,4-substituents in a syn orientation.

Fig. 8 Radial hexagons for various synthesis plans for (+)-sertraline hydrochloride.
Fig. 9 Kernel waste distribution profiles for various synthesis plans for (+)-sertraline hydrochloride.
Fig. 10 Total waste distribution profiles for various synthesis plans for (+)-sertraline hydrochloride.
Fig. 11 E-factor profiles for various synthesis plans for (+)-sertraline hydrochloride: (A) ranking according to kernel mass of waste produced, (B) ranking according to year of reporting of synthesis plan. The plan-year correspondences are as follows: Pfizer G1 (1984), Lautens (1999a), Pfizer G2 (1999b), Buchwald (2000), Gedeon Richter (2002), Pfizer G3 (2004), Lautens G2 (2005), Zhao (2006), Backvall (2010).
Fig. 12 Target bond maps for various synthesis plans (+)-sertraline hydrochloride.
Fig. 13 Ring construction strategies employed for various syntheses of (+)-sertraline hydrochloride.
CONCLUDING REMARKS

The results of the extensive synthesis database reveal the following summary of effective green synthesis strategies:

- minimizing overall number of steps required
- maximizing number of target bond-forming reactions
- minimizing waste contribution from sacrificial reactions
- making use of “aufbau”-type reactions: multicomponent reactions (MCRs), tandem, domino, cascade, and “one-pot” reactions as central theme reactions in synthesis planning especially for the construction of ring systems
- making use of additive type redox reagents where hydrogen or oxygen atoms are added to the molecular framework and eventually end up in the final target product structure
- reducing the degree of oscillation in oxidation level of atoms in target bond-forming steps over the course of the synthesis plan—better plans are those that involve steady increases or decreases in oxidation level
- telescoping or concatenation of steps
- reducing solvent demand
- choosing safe and benign methodologies to carry out transformations

With over 12,000 reactions examined in the literature for the construction of the synthesis database, it was clearly apparent that there are serious problems with the degree of correctness and detail given in literature reports, particularly in experimental sections of publications. The reviewing of experimental sections is at best casual especially that it is now customary to include them as part of the supplementary or supporting materials rather than as part of the main body of the papers. A fair assessment of greenness by any kind of metrics analysis necessitates the full disclosure of all masses of materials used in a chemical process in addition to the mass of the final target product. Such gaps obviously weaken arguments that favor one plan or process over another. More broadly, this situation jeopardizes the legitimization and implementation of green chemistry principles if proper assessments cannot be made because of limited availability of essential data. The following problems are particularly troublesome:

- the use of phrases such as “in our hands”, “readily available”, “obtained in good yield”, and “worked up in the usual way”
- missing information in experimental procedures such as: reaction yields for all steps, especially in communications; masses of gaseous reagents; masses of materials used in work-up procedures such as volumes of wash or extraction solvents and masses of drying agents; and masses of materials used in purification procedures such as mass of silica gel and volume of solvent eluents used in column chromatography, and volumes of recrystallization solvents
- no reporting of energy consumption as part of the standard protocol for any chemical reaction
- synthesis of specialized catalysts, often having more complicated structures than the substrates on which they act, are not mentioned or included in the overall analysis of material efficiencies of plans
- errors in the correspondence between masses and moles of materials used and in reaction yield calculations, particularly for steps involving classical resolutions
- disguising poor reaction yield performance (less than 50 %) by reporting percent selectivity, percent conversion, or percent yield based on recovered starting material, rather than by reporting the actual value based on the limiting reagent used
- exaggeration of claims of “greenness” or “conciseness” by highlighting improvements in one or two parameters, such as percent overall yield and number of steps, without considering a more global approach that involves a suite of metric performance variables.
Given these problems and the recognized urgency in implementing green chemistry in the mainstream practice of synthetic organic chemistry, some guidelines for standardization and ranking of plans are suggested. These include:

- disclosure of all reaction parameters in a given plan including both material and energy consumption
- mandatory reporting of green metrics as part of the standard protocol in reports of synthesis plans in the literature, especially if those plans are advertised as “greener” than prior published ones for the same target molecule
- agreeing to a common set of feedstocks that all synthesis plans to a given target can be compared and ranked—this task can be initiated by an international committee composed of industrial and academic chemists struck under the auspices of IUPAC
- changing the culture of casual reviewing of experimental protocols
- implementation of metrics calculations, such as the radial pentagon analysis, as a standard proof-reading tool offered by editors of journals for authors to check their own work as part of the submission process for their manuscripts
- initiation of a Green Syntheses series with the same high standards and purpose as the venerable series Organic Syntheses launched by Roger Adams in the 1930s

REFERENCES