Reaction Yields

Deutsche Ausgabe: DOI: 10.1002/ange.201503890
Internationale Ausgabe: DOI: 10.1002/anie.201503890

A Priori Estimation of Organic Reaction Yields**

*Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga, and Bartosz A. Grzybowski**

Abstract: *A thermodynamically guided calculation of free energies of substrate and product molecules allows for the estimation of the yields of organic reactions. The non-ideality of the system and the solvent effects are taken into account through the activity coefficients calculated at the molecular level by perturbed-chain statistical associating fluid theory (PC-SAFT). The model is iteratively trained using a diverse set of reactions with yields that have been reported previously. This trained model can then estimate a priori the yields of reactions not included in the training set with an accuracy of ca. ±15%. This ability has the potential to translate into significant economic savings through the selection and then execution of only those reactions that can proceed in good yields.*

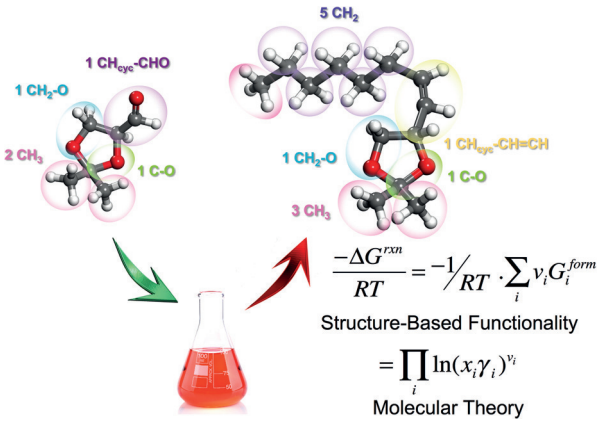


Figure 1. Representation illustrating the decomposition of substrate and product molecules into functional groups, calculation of free energy from group contributions, and of the activity coefficients from PC-SAFT molecular theory.

Performing a reaction and laborious work-up only to discover a few-percent yield is probably one of the most frustrating experiences of the chemical profession. In this spirit, the ability to estimate the yields of organic reactions before they are actually performed could translate into immense economical (and environmental) savings as it would guide the chemists to perform only those reactions that have a realistic chance to proceed in decent yields. Since most reactions in organic chemistry are under thermodynamic control,[1] the principles of thermodynamics appear a suitable starting point for analyses in which reaction yields would be related to appropriately defined and optimized reaction free energies ΔG (Figure 1a). While simple in concept, precise calculation of free energies of structurally diverse organic molecules is a rather difficult task since even subtle variations in the molecular structure may drastically change the Gibbs free energy of molecule formation. Additionally, any realistic estimates have to consider non-idealities of a system[2] at a particular concentration, temperature, and pressure; these non-idealities are often embodied in the activity coefficients, which unfortunately are difficult to determine experimentally and are often overlooked. It follows that any generally applicable yield calculations must rely on molecular-level theories that need to be carefully adapted (or even extended) to account for structural diversity of organic molecules. In light of these difficulties it is perhaps not surprising that to date there has been only few attempts to attack the problem; for example, Hoffmann et al.[3] calculated the activity-based reaction constants over a narrow set of reactions while Hukkerikar and co-workers used structure-based correlations to estimate free energies of individual small organic molecules[4] but not of complete chemical reactions. Most importantly, there has been no example reported of a model that would not only correlate experimental and calculated yields for a training set of reactions, but would then also validate the predictions of the model for other test compounds not included in the original training set.

This is precisely what we attempted in the current work by i) calculating free energies of formation of compounds by assuming additive contributions[4] of their constituent fragments/functional groups; ii) applying the molecular-level perturbed-chain statistical associating fluid theory (PC-SAFT)[5] to incorporate the activity coefficients and equilibrium concentrations; and iii) optimizing the group contribu-

[*] Dr. F. S. Emami,[+] Dr. A. Vahid,[+] E. K. Wylie
Department of Chemical and Biological Engineering
Northwestern University (USA)
S. Szymkuć, Dr. P. Dittwald, K. Molga, Prof. B. A. Grzybowski
Institute of Organic Chemistry, Polish Academy of Sciences
Warsaw (Poland)
Prof. B. A. Grzybowski
Department of Chemistry and the IBS Center for Soft and Living Matter, UNIST, Ulsan (South Korea)
E-mail: grzybor72@unist.ac.kr
[+] These authors contributed equally to this work.

[**] This work was supported by 1) NERC, which was an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0000989 and 2) the Symfonia Award, grant number 2014/12/W/ST5/00592 from the Polish National Science Center, NCN. B.A.G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1. The authors also thank ProChimia Surfaces (Gdansk, Poland) for providing the dataset of reactions performed multiple times (Supporting Information, Section S9).

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201503890.

tions on a diverse training set of 23 000 reactions for which the yields and full stoichiometries have been reported. We then tested the predictive power of the model thus built against test sets of other reactions (again, structurally diverse) achieving the accuracy of yield prediction to within 15%. While the model rests on several simplifying assumptions (which we emphasize below), the level of accuracy is already practically relevant as it allows for the statistically significant a priori discrimination between poor-, average-, and high-yielding reactions.

The underlying thermodynamic principle is illustrated in Figure 1 and has two parts to it. On one hand, substrate and product molecules (species indexed *i*) of a given reaction are partitioned into smaller fragments/functional groups for which Gibbs free energies of formation, G_i^{form}, are calculated. The summation over these free energies with appropriate stoichiometric coefficients (ν_i) then gives ΔG_{calc}. These values, however, are not yet accurate since they do not account for any non-idealities of the system (for example, solvent effects); consequently, they constitute only our initial guess values which need to be further optimized. This we do by training the model based on previously reported yields. From these experimentally recorded yields and from the activity coefficients calculated based on the molecular-scale PC-SAFT theory, and correcting for solvent effects, we back-track the experimental reaction free energies, ΔG_{exp}. We then compare the values ΔG_{calc} and ΔG_{exp} and iteratively adjust the values of group contributions G_i^{form} until convergence (that is, until the correlation between ΔG_{calc} and ΔG_{exp} is maximized). The values of group contributions, G_i^{form}, optimized in this way are then used to make predictions of free energies and yields of other reactions, not included in the training set. All of the theoretical details included in the Supporting Information (Sections S2–S6); the key steps of the above procedure are as follows.

1) Choice of test set. As a training set, we chose and manually curated[6] a set of 23 000 previously reported reactions for which reaction conditions, stoichiometries, and experimental yields were all available. The set was chosen at random with the proviso that it contained both small and large molecules (MW between 100 and 1000) in proportions similar to those in the entire body of known organic reactions[7] (Supporting Information, Figure S10). The set was structurally diverse as evidenced by the pairwise Tanimoto coefficient[8] similarity map of reaction substrates and products (Supporting Information, Figure S11) and by clustering analysis (Figure S12).

2) Decomposition into groups. All participating molecules were decomposed into 296 functional groups listed in the Supporting Information, Table S2 for which the initial (guess) values of Gibbs free energies of formation at 298 K, $G_i^{\mathrm{form,298\,K}}$, were taken from Ref. [4] (Supporting Information, Table S3). The decomposition procedure was hierarchical in the sense that functional groups were matched against the molecule of interest in the descending order of their complexity (that is, more complex groups were matched first.[9]

3) Calculation of initial-guess reaction free energies. Thereafter, free energies of formation at 298 K of all substrates and products were calculated by summing-up group contributions, and the reaction free energies were obtained from $\Delta G_{\mathrm{calc}} = \Sigma_i \nu_i G_i^{\mathrm{form,298\,K}}$, where ν_i are the stoichiometric coefficients. These initial-guess values were further corrected to the previously reported temperatures using heat capacities, enthalpies, and thermodynamic relationships (Supporting Information, Section S6 and Table S4).

4) Independent calculation of reaction free energies from experimental yields. As we already mentioned, the values of $\nu_i G_i^{\mathrm{form}}$ and ΔG_{calc} calculated without considering any non-idealities were inaccurate and, as we verified, could not be used to predict previously reported yields. Consequently, we trained our model against experimental yields and the non-idealities/solvent effects they entailed. To do so, we first converted the experimental yields, ξ, to molar fractions, x_i, via $\xi = (n_i^0 - x_i n^0)/(x_i \nu - \nu_i)$, where n_i^0 stands for the initial number of moles of substance *i*, n^0 denotes the total number of initial moles, ν_i are stoichiometric coefficients, and ν is the so-called total stoichiometry coefficient defined as $\nu = \Sigma_i \nu_i$. Next, to account for non-idealities, we calculated activity coefficients γ_i of all substances. These calculations were based on the well-validated PC-SAFT molecular theory[10,11] in which free energies of molecules comprise terms dues to hard-chain interactions, dispersion attractions, and short-range hydrogen bonding (Supporting Information, Equations S6–S24). Accounting for solute–solvent interactions but not for solute–solute ones (that is, taking infinitely dilute solutes in liquid solvent as the reference points for substances), the free energies could be converted into fugacity coefficients (Supporting Information, Sections S2–S4). We note that the high-dilution assumption simplified the calculations immensely (indeed, we found that otherwise they would be computationally prohibitive owing to extensive multicomponent phase equilibria calculations for every single reaction in our database) while being realistic for typical organic reactions (for example, for 1 mM concentrations, the average distance between the molecules is on the order of 12 nm; for 100 mM it is on the order of 2 nm). Finally, the activity coefficient of every species *i* was calculated as a ratio of the fugacity coefficients, φ, of the pure solvent and the dilute (non-interacting) solute $\gamma_i = \varphi_{\mathrm{pure\,solvent}}/\varphi_{\mathrm{dilute\,solute}}$.

With the details of all these lengthy calculations relegated to the Supporting Information (Sections S2–S4), the key thing to note is that with the knowledge of both the composition of the reaction mixture x_i and the activity coefficients γ_i we can use the mass action formula to derive reaction free energies as $\Delta G_{\mathrm{exp}} = -RT \ln \prod (x_i \gamma_i)^{\nu_i}$. We emphasize that using the activity coefficients is essential for this approach to be meaningful and accurate; a naïve expression based solely on the mole fractions (that is, $\Delta G_{\mathrm{exp}} = -RT \ln \prod (x_i)^{\nu_i}$), would be valid only for ideal solutions at standard conditions.

5) Optimization of group free energies. The last step of the analysis was to optimize the values of group contributions G_i^{form}. To this end, an objective function was defined as $\mathrm{OBJ} = \sqrt{\sum_{j=1}^{n} (\Delta G_{\mathrm{exp}} - \Delta G_{\mathrm{calc}})^2 / n}$, where $n = 23\,000$ is the number of reactions in the training set. Free energies of groups G_i^{form} were optimized iteratively to minimize the OBJ by applying the globally convergent Newton optimization technique.[12] Each iteration involved changing the values of G_i^{form} until convergence of OBJ below a desired threshold was achieved.

At the same time, the activity coefficients were calculated only once at the beginning of the routine, since their values do not change in each iteration.

The above optimization procedure was implemented on a Northwestern's Quest computer cluster using a single processor and convergence was achieved after about four weeks of calculation. The results in Figure 2a show that the

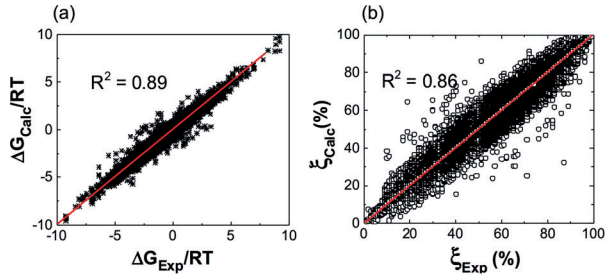


Figure 2. Correlations between experimental and calculated a) reaction free energies and b) reaction yields. Correlations were optimized for the $n = 23\,000$ reactions from the training set.

optimized values of G_i^{form} and the related ΔG_{calc} values correlated against the previously reported derived values with R^2 coefficient close to 0.9. Moreover, when the free calculated free energies were converted to reaction yields, the correlation was still very high, with $R^2 = 0.86$ (Figure 2b).

Three remarks are in order. First, it is important to note that the optimization procedure involving 296 free parameters (that is, G_i^{form} values of groups) is not necessarily guaranteed to converge. In this light, our choice of the initial guess values is a judicious one, as it gives robust convergence to desired threshold levels. Even though the existence of other initial parameter sets that would also give convergence cannot be excluded with certainty, we were not able (despite multiple tries) to identify such sets. In particular, the initial guess taken without calculating activity coefficients and correcting for non-idealities did not lead to a predictive model, again emphasizing the importance of the thermodynamic basis of our approach (rather than it being just a multidimensional parameter optimization).

Second, in our approach we converted experimental yields into free energies and then optimized these values against free energies calculated from first principles. Another approach could be imagined in which the calculated guess free energies are first converted into yields that are then optimized against experimental yield values. However, because yields are non-linear (exponential) functions of reaction free energies, the optimization procedure is significantly more challenging: in fact, we were not able to achieve satisfactory convergence of calculated versus experimental yields.

Third, inspection of specific molecules indicated that largest errors in the analysis stem from imperfect decomposition of molecules into functional groups. One possibility here is that the molecules include motifs that are not in the list of 296 groups. Naturally, one could increase the number of groups; however, accounting for all types of, for example, exotic heterocycles would bring the number to several thousand groups for which multivariate optimization becomes unrealistic. The choice of 296 groups yielding correlation coefficients close to 0.9 appears a viable tradeoff between the computational requirements and the accuracy of the method. Another important source of errors is in cases when the algorithm incorporates bonds that are actually made/broken in a reaction to two different groups (that is, bond assigned to one group in reactants, but to another group in products). In this way, the bond formation free energies are not properly accounted for and the yields are nonsensical: we observe this problem in about 10% of the molecules, which were eliminated from the dataset.

So far, we have shown that calculation parameters can be successfully optimized to reproduce the previously reported yields for training set reactions. For any practical relevance, however, our model needs to be able to predict (without any further training/optimization) the yields of reactions not included in the training set. To do so, we have chosen at random additional $m = 3000$ previously reported reactions with known yields. This test set was structurally diverse, as evidenced by the pairwise similarity maps such as that in Figure 3a (see also the Supporting Information, Figure S2a,b). The key result in Figure 3b shows the map correlating literature and calculated yields for the entire test set: there is a clear trend along the diagonal with Pearson correlation coefficient[13] of 0.68. Also, Figure 3c plots the means and standard deviations of the absolute differences between predicted and experimental yields $|\xi_{\text{pred}} - \xi_{\text{exp}}|$. To detect any potential systematic biases and determine asymp-

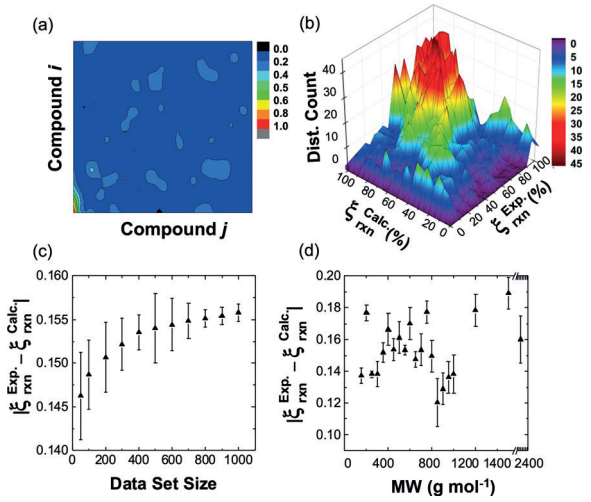


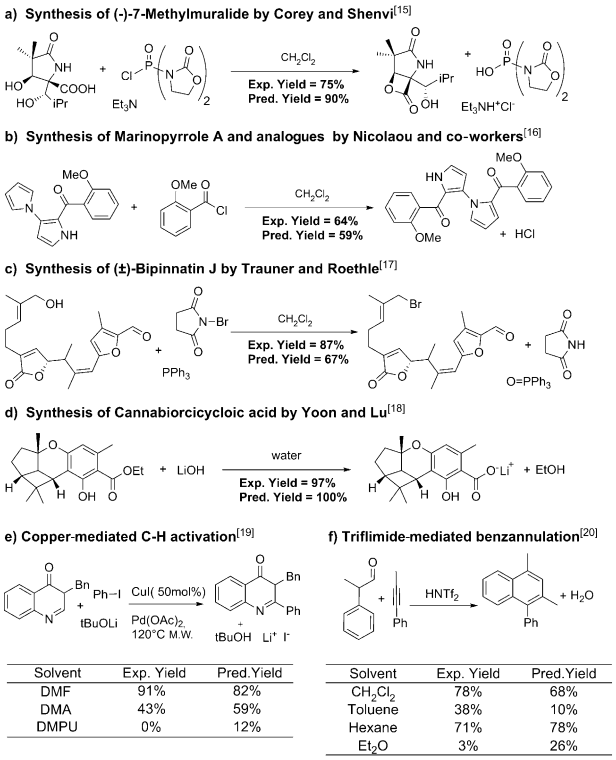
Figure 3. Prediction of reaction yields over the test set and its subsets. a) A map of pairwise Tanimoto coefficients calculated for the primary products of the $m = 3000$ reactions indicates that the test set is structurally diverse (blue color corresponds to low similarity; see also the Supporting Information, Figure S2). b) Histogram of the numbers of reactions characterized by given values of calculated and experimental yields ($\xi_{\text{exp}}, \xi_{\text{pred}}$). Data is scattered within about 15% of the (0,0) to (100,100) diagonal corresponding to perfect prediction. c) Average values and standard deviations of $|\xi_{\text{pred}} - \xi_{\text{exp}}|$ calculated for various-size subsets of the test set. d) The average values and standard deviations of $|\xi_{\text{pred}} - \xi_{\text{exp}}|$ for molecules of different molecular weights. The accuracy of the model does not depend on MW.

totic behavior, this analysis was performed on various sub-sets $\{m_i\}$ of the entire test set $\{m\}$. Specifically, we chose eleven sizes of the smaller subsets $\{m_i\} = 50,100,200,300,\ldots1000$. For each size, we generated ten samples (chosen randomly from amongst the entire $m = 3000$ molecules of the test set). We then calculated $|\xi_{pred} - \xi_{exp}|$ for each set and then took average and standard deviation for each m_i. The results evidence that as the size of the test set increases, the difference between calculated and experimental yields asymptotically approaches about 15% with statistical standard deviation of $\pm 2\%$. Moreover, in Figure 3d we summarize calculations in which we analyzed the values $|\xi_{pred} - \xi_{exp}|$ as a function of molecular weights of the reaction products; as seen, there is no systematic trend. Together, these findings show that our model predicts reaction yields with an accuracy of $\pm 15\%$ irrespective of the mass and complexity of the reacting molecules.

The latter point is further illustrated in Figure 4a–d, which compares predicted vs. experimental yields of reactions selected from some classic syntheses leading to relatively complex natural products. A related and important issue is the performance of the model in the light of inherent variability of literature reported yields. Here, we examined two additional test sets. The first was based on a sample of 200 reactions from *Organic Syntheses* (all listed in the Supporting Information, Section S8), which were run on relatively large scales and independently checked by a second laboratory. The quality of the algorithm predictions (Supporting Information, Figure S14) was about $16.3 \pm 2\%$ in terms of the average absolute error; that is, similar to the error for the 3000 Reaxys test set. In the second case, we examined reactions performed in industry over many years in independent batches; this particular dataset was generously provided to us by ProChimia Surfaces and is listed in the Supporting Information, Section S9. Even though the experimental yields had an inherent spread (± 3–15% from the average yield), the theoretical predictions still matched the average experimental values to within $15.7 \pm 11\%$ absolute error. Finally, we considered whether the predictions of the algorithm reflect not only the structures of the substrates/products but also reaction conditions. For this case, Figure 4e,f have two illustrative recently reported examples, where the reaction yields were found to vary drastically upon often subtle changes in the structure of the solvent molecules; again, the theoretical predictions reproduce the experimental trends.

The last issue we wish to signal is the scalability of the predicted yields to large-scale reactions. In principle, all the considerations of the present work are independent of scale, since reaction free energies per mol (from which the yields are derived) are intensive variables. In practice, however, the yield is affected by the design of the reaction vessel (for example, quality of mixing in a reactor) and, above all, by the fact that under scale-up conditions, the Gibbs free energy of reaction is affected by heat loss (dissipation).[21] Theoretical discussion of such effects along with an illustrative example of an industrially relevant reaction is provided in the Supporting Information, Section S7.

In summary, the model we developed is, to the best of our knowledge, the first successful attempt to estimate the yields of organic reactions with any practically relevant accuracy. Even though the model does not offer single-digit accuracy (which is unrealistic given the variability of experimental data on which it is trained), it can robustly distinguish between poor, average, and high-yielding reactions with statistical significance. This capability can be relevant to synthetic planning, including computer-assisted synthesis,[14] in which a chemist is often offered large numbers of synthetic choices from which only few can be executed.



Solvent	Exp. Yield	Pred.Yield
DMF	91%	82%
DMA	43%	59%
DMPU	0%	12%

Solvent	Exp. Yield	Pred.Yield
CH_2Cl_2	78%	68%
Toluene	38%	10%
Hexane	71%	78%
Et_2O	3%	26%

Figure 4. a)–d) Predicted vs. experimental yields for some examples taken from the "classic" organic literature.[15–18] e), f) Two examples from recent reports where the model captures pronounced yield differences in different solvents.[19,20] M.W. = microwave.

Keywords: optimization · reaction yields · thermodynamics

How to cite: *Angew. Chem. Int. Ed.* **2015**, *54*, 10797–10801
Angew. Chem. **2015**, *127*, 10947–10951

[1] While some classes of reactions (Wittig, Diels–Alder, enolizations) have a well-documented propensity to give kinetic products, the numbers of such cases are low compared to the numbers of reactions that proceed under thermodynamic control. For the statistical analysis of thermodynamically vs. kinetically controlled reactions, please see the Supporting Information, Section S1.1.
[2] R. A. Heidemann, J. M. Prausnitz, *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 1773–1776.
[3] P. Hoffmann, M. Voges, C. Held, G. Sadowski, *Biophys. Chem.* **2013**, *173*, 21–30.
[4] A. S. Hukkerikar, B. Sarup, A. Ten Kate, J. Abildskov, G. Sin, R. Gani, *Fluid Phase Equilib.* **2012**, *321*, 25–43.
[5] J. Gross, G. Sadowski, *Ind. Eng. Chem. Res.* **2001**, *40*, 1244–1260.

[6] The reactions were all either from published scientific papers or patents stored in the Reaxys database. As described in our previous works on the network of organic chemistry (see Refs. [7,14]), all reactions in our dataset were first analyzed by house-written scripts to ensure that our dataset did not contain any duplicate entries that would bias the optimization analyses. All reactions remaining in the dataset were curated/scrutinized to ensure atomic balances. In some cases, the missing simple molecules (such as water, ethanol) could be added by house-written scripts. In other cases, source literature had to be consulted.
[7] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263 – 7269; *Angew. Chem.* **2005**, *117*, 7429 – 7435; K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2006**, *45*, 5348 – 5354; *Angew. Chem.* **2006**, *118*, 5474 – 5480; A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2014**, *53*, 8108 – 8112; *Angew. Chem.* **2014**, *126*, 8246 – 8250.
[8] D. J. Rogers, T. T. Tanimoto, *Science* **1960**, *132*, 1115 – 1118.
[9] S. Soh, Y. Wei, B. Kowalczyk, C. M. Gothard, B. Baytekin, N. Gothard, B. A. Grzybowski, *Chem. Sci.* **2012**, *3*, 1497 – 1502.
[10] M. Kleiner, J. Gross, *AIChE J.* **2006**, *52*, 1951 – 1961.
[11] J. Vrabec, J. Gross, *J. Phys. Chem. B* **2008**, *112*, 51 – 60.
[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York, **2007**.
[13] S. M. Stiglèr, *Stat. Sci.* **1989**, *4*, 73 – 86.
[14] a) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928 – 7932; *Angew. Chem.* **2012**, *124*, 8052 – 8056; b) C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. H. Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7922 – 7927; *Angew. Chem.* **2012**, *124*, 8046 – 8051; c) M. Peplow, *Nature* **2014**, *512*, 20 – 22.
[15] R. A. Shenvi, E. J. Corey, *J. Am. Chem. Soc.* **2009**, *131*, 5746 – 5747.
[16] K. C. Nicolaou, N. L. Simmons, J. S. Chen, N. M. Haste, V. Nizet, *Tetrahedron Lett.* **2011**, *52*, 2041 – 2043.
[17] P. A. Roethle, D. Trauner, *Org. Lett.* **2006**, *8*, 345 – 347.
[18] Z. Lu, T. P. Yoon, *Angew. Chem. Int. Ed.* **2012**, *51*, 10329 – 10332; *Angew. Chem.* **2012**, *124*, 10475 – 10478.
[19] S. Laclef, M. Harari, J. Godeau, I. Schmitz-Afonso, L. Bischoff, C. Hoarau, V. Levacher, C. Fruit, T. Besson, *Org. Lett.* **2015**, *17*, 1700 – 1703.
[20] S. Ponra, M. R. Vitale, V. Michelet, V. Ratovelomanana-Vidal, *J. Org. Chem.* **2015**, *80*, 3250 – 3257.
[21] M. D. Koretsky, *Engineering and Chemical Thermodynamics*, 2nd ed., Wiley, Hoboken, **2013**.

Received: April 28, 2015
Published online: July 21, 2015

