Electronic Supplementary Material (ESI) for Organic & Biomolecular Chemistry.

Supporting Information

Current Complexity: A Tool for Assessing the Complexity of Organic Molecules

Jun Li and Martin D. Eastgate*

Chemical Development, Bristol-Myers Squibb, 1 Squibb Drive, New Brunswick, NJ, 08903 (USA)

Table of Contents

A training dataset of 40 molecules from both BMS internal programs and literature compounds (natural products and man-made systems) were carefully selected for diversity in structural class, topology, substitution pattern, presence/absence of stereochemistry and functional groupings. Distributions of various molecular properties from the selected group are illustrated as follows:

PageS3

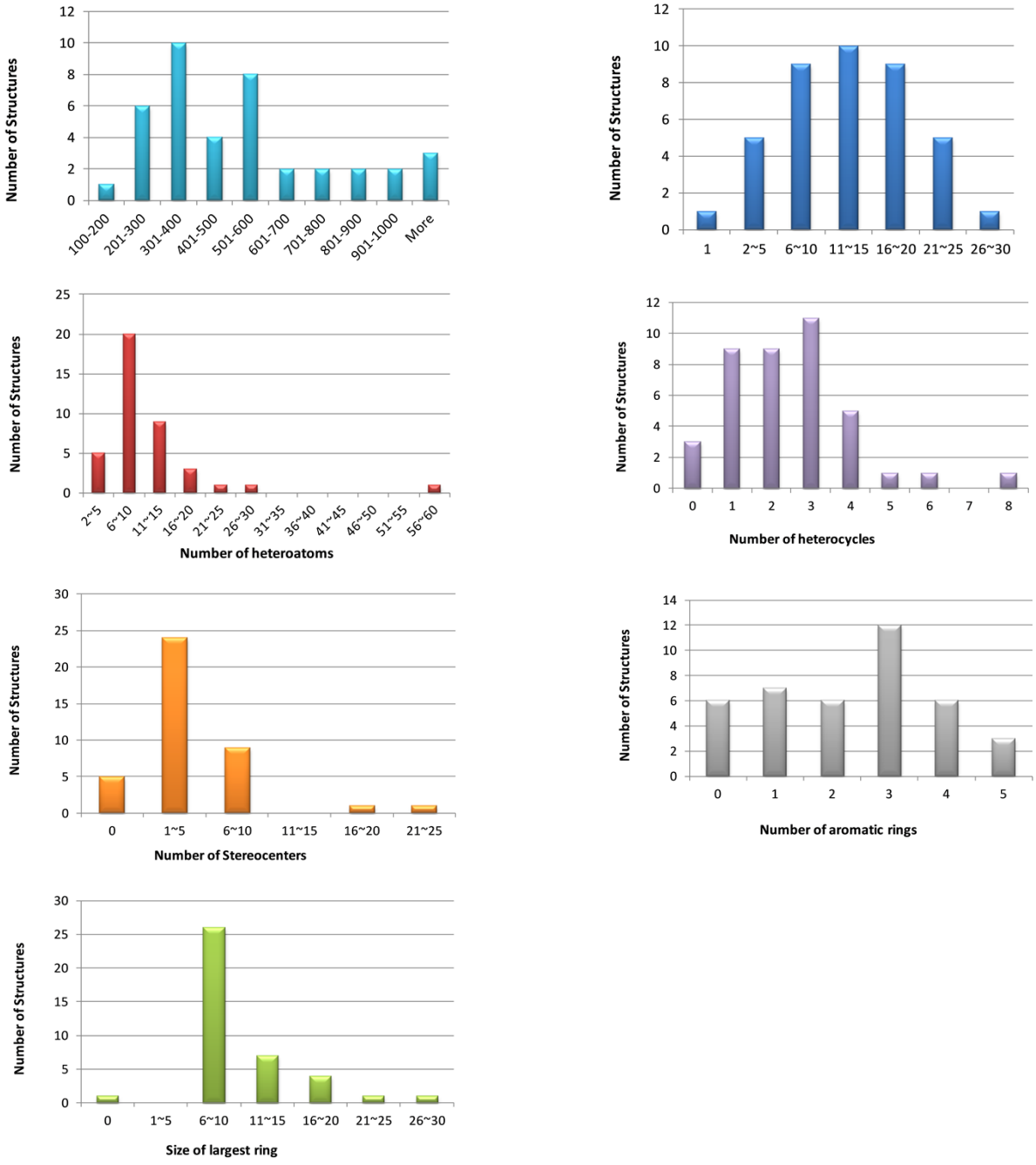


Figure S1 Distributions of various molecular properties from MW, double bond equivalent (DBE), heteroatoms, heterocycles, aromatic rings, stereocenters, and size of largest ring in the training dataset.

Collated training set structures were provided to a group of highly trained synthetic organic chemists within BMS, with an average 10 to 20 years experience in synthetic organic chemistry from both academic and industrial settings. In our initial assessment, no synthetic information was given to the chemists – they could look on their own, but the information was not supplied. To a subset of our raters we gave *all* the relevant synthetic information. The group was asked to force rank the molecules from 1 to 40 with 1 being the most complex. The correlation chart summarized the pooled data with comparisons between raters; on the right of the diagonal are the pair-wise correlations, with red stars signifying significance levels. The more red stars the higher the correlation. As the correlations get bigger, the font size of the coefficient increases. On the left side of the diagonal is the scatter-plot matrix between two chemists, with loess fitting line in red to illustrate the underlying relationship. In this chart, lots of correlations were observed, indicating an overall agreement among the raters.

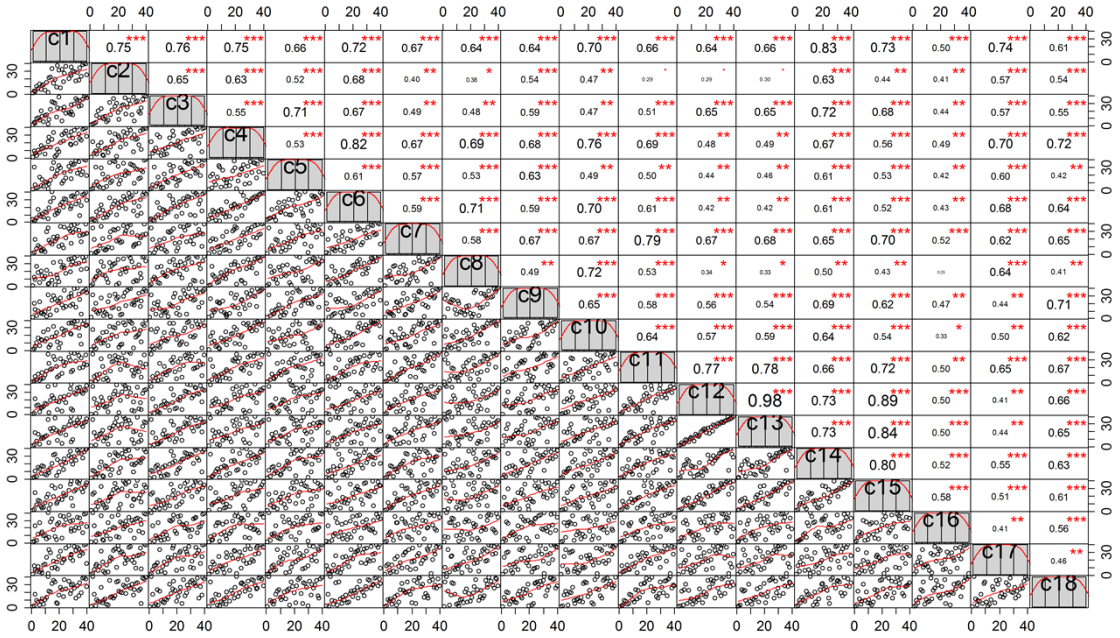


Figure S2 Force ranking pair-wise correlation chart of chemists.

To quantify structural information of molecular networks, we leveraged a quantitative network analysis tool, which allows us to access a large number of topological indices.[1] Since the topological indices do not contain heteroatoms and stereo centers those factors were taken into account separately. For stereogenic centers, in order to differentiate the contributions from either *de novo* synthesis, or naturally occurring sources, we counted the number of stereocenters made during the synthesis (chiral_made) separately from those purchased. For aryl heterocycles, we understand that unpredicted stereo-electronic effects as well as other latent complexity in terms of the proximal disposition of heteroatoms present could be challenging for prediction. We designated 'HAA' as the total number of heteroatoms on and in aromatic rings, separate from those on aliphatic parts of the molecule. From fragment-based approaches, we also investigated the use of heterocycle fragment prevalence to gauge synthetic complexity assuming the fragment frequency is related to synthetic accessibility. In terms of other intrinsic complexity factors, we used 'HANR' as heteroatoms on nonaromatic moieties and 'DSC' as well as 'DSH' to define number of unsaturations in aliphatic and aromatic moieties, respectively. For extrinsic (variable) complexity, we chose to follow step counts, ideality from Baran's pioneering work, and yield. Given those postulated factors, we applied a regression subset selection approach (LEAPS) [2] to probe the combinations of main factors as well as two-way interactions of the factors. Selected factors were then evaluated in a following Bayesian ordinal probit model for the ordinal scale between 1 to 10. Topological factors such as Randic, Zagreb, and Estrada indices share similar trends, we chose Randic to represent the molecular structural information. Factors such as unsaturation also played a role in contributing to the chemist ranking, however, it was not useful in the model. Both HNAR and aryl heterocycle prevalence did not give significant contributions in the Bayesian model. Considering variability of yield reported in the literature and inherent inaccuracy of yield information, we chose to not include yield in the final analysis. To differentiate convergent and linear synthesis, we count longest linear steps and add 50% of all the remaining branching steps, this was a valuable differentiation in the model. Eventually, we found that five major factors associated with the complexity ranking score: i) a molecular topological index (Randic); ii) number of stereogenic centers

established via *de novo* synthesis; iii) heteroatoms on and in aromatic rings; iv) step counts associated with the synthesis; v) percent ideality of the current route.

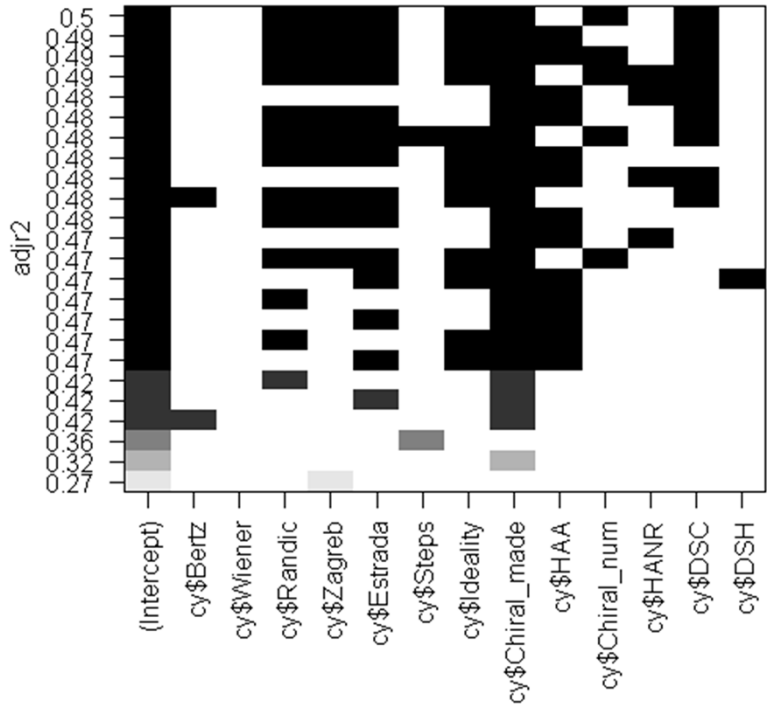
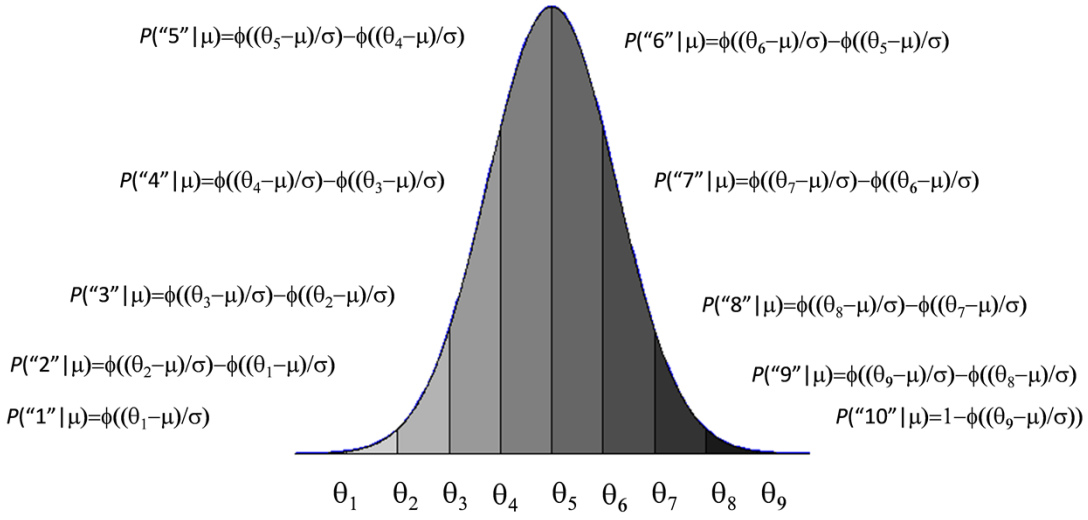


Figure S3 Example of the regression subset approach showing linear regression fitting with adjusted R squared.

Given the eccentric nature of human assessments, probabilistic modelling has been highly recommended for human rating systems. In order to link the ordinal 1 to10 ranking values from the chemists to the postulated underlying factors, an ordinal probit regression was established where the linear regression output was mapped to the ordinal value depending on which threshold, θ, falls between via a cumulative normal density function, ϕ. The space between threshold θ is not required to be evenly divided. Bayesian inference of this regression model essentially reallocates the credibility across the space of model parameters to be consistent with observed chemists ranking data. This model generates the probability of a molecule being in each one of the 10 groups of complexity.



$$\mu=\beta_0+\beta_{Randic}X_{Randic}+\beta_{ss}X_{ss}+\beta_{HAA}X_{HAA}+\beta_{Steps}X_{Steps}+\beta_{Ideality}X_{Ideality}+\varepsilon$$
$$\varepsilon=N(0,\sigma^2)$$

Figure S4 Underlying mapping in ordinal probit regression model.

Bayesian ordinal probit regression analysis were carried out using either WinBUGS or openBUGS[3] and analyzed with R 2.15.2 statistical programming language[4] or MCMCpack R package[5]. Both approaches provide equivalent results. Typically, three chains, 2000 burn-in steps, and 5000 steps for each chain were used in this model without

issues in convergence. The following Monte Carlo (MC) diagnostics were generated from posterior distribution using openBUGS.[6] The MC convergence was confirmed by checking well-mixed chains and analysis of the variance within- and between-chains using Brooks-Gelman-Rubin (BGR) diagnostic.

MCMC Diagnostics:

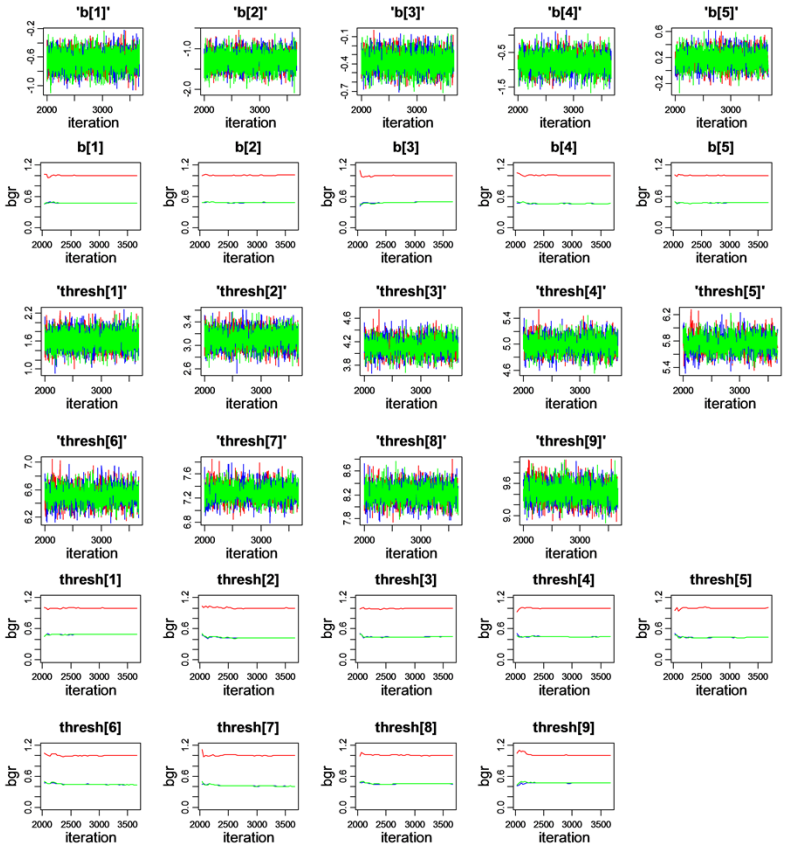


Figure S5 Plots of well-mixed MC chains and BGR plots for each factor coefficient and threshold.

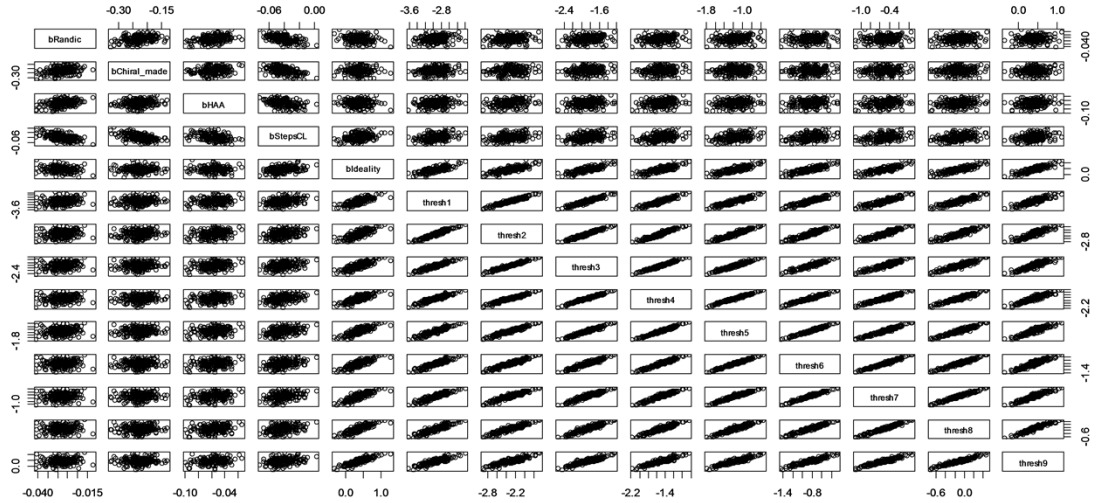


Figure S6 Pair-wise scatter plots of each factor coefficient and threshold.

The credible values of the regression coefficients and thresholds are shown in the following histograms. Each histogram is marked with its 95% highest density interval (HDI), which summarizes where the bulk of the most credible values fall. It should be noted that for threshold θ distributions, they are highly correlated from scatter plot and the overlap between parameter distributions of θ_i and θ_{i+1} does not violate the ordering of the thresholds. The actual differences θ_i-θ_{i+1} showed no violation of the ordering in threshold.[7]

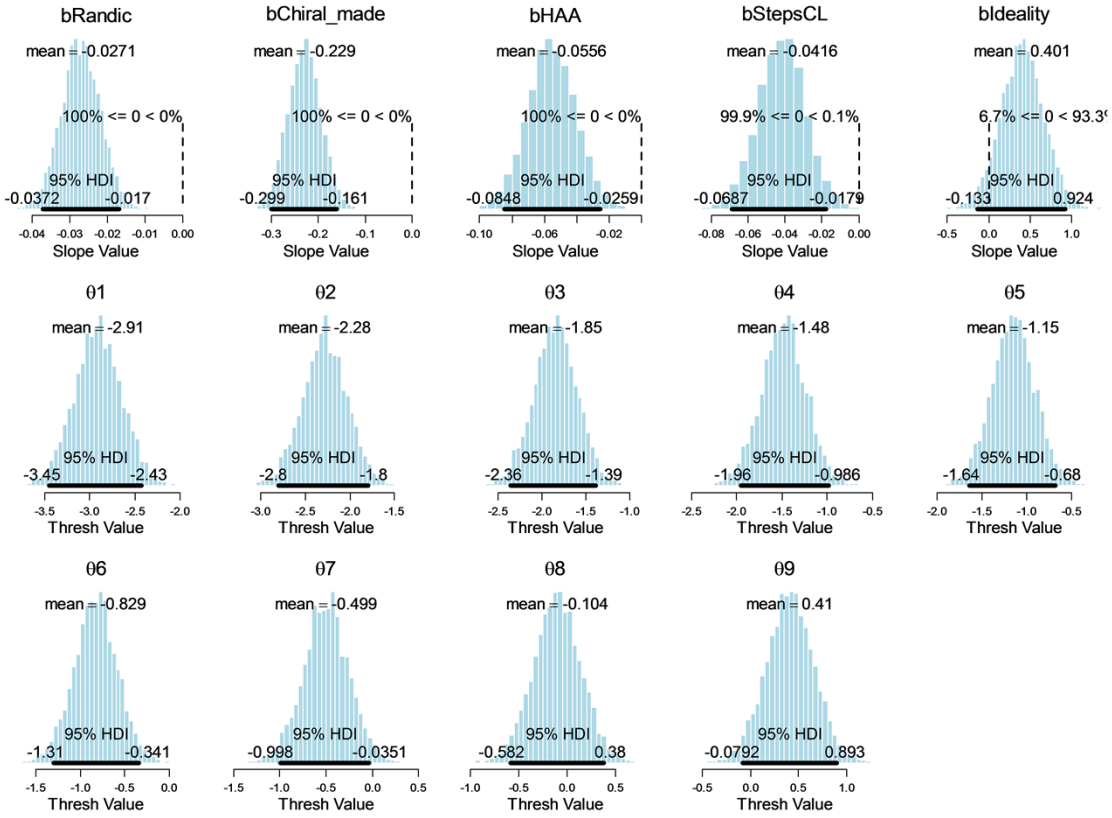


Figure S7 Histograms of each factor coefficient slope and threshold.

Correlation coefficient between weighted prediction index and weighted chemist ranking score is 0.84, which is reasonable compared to other models in the literature,[7] signifying the validity of the predictive model. We have conducted assessment over a validation set of many (>60) BMS internal molecules and found that this approach passes the 'common-sense' check, with molecules falling into the correct locations on the indexed scale. Due to the highly complex and eccentric nature of the human rating, we found this probability model captured the majority of the consensus among the raters.

Expanding the training set of molecules to include much less complex systems and increase the pool of experienced chemists will help adjust some of the deficiencies in the model – this is on-going work. Furthermore, other factors such as chemical reactivity, thermal stability, cost, and physical characteristics associated with specific intermediates and reagents could be included if process safety, economy, and ease of purification are taken into account.

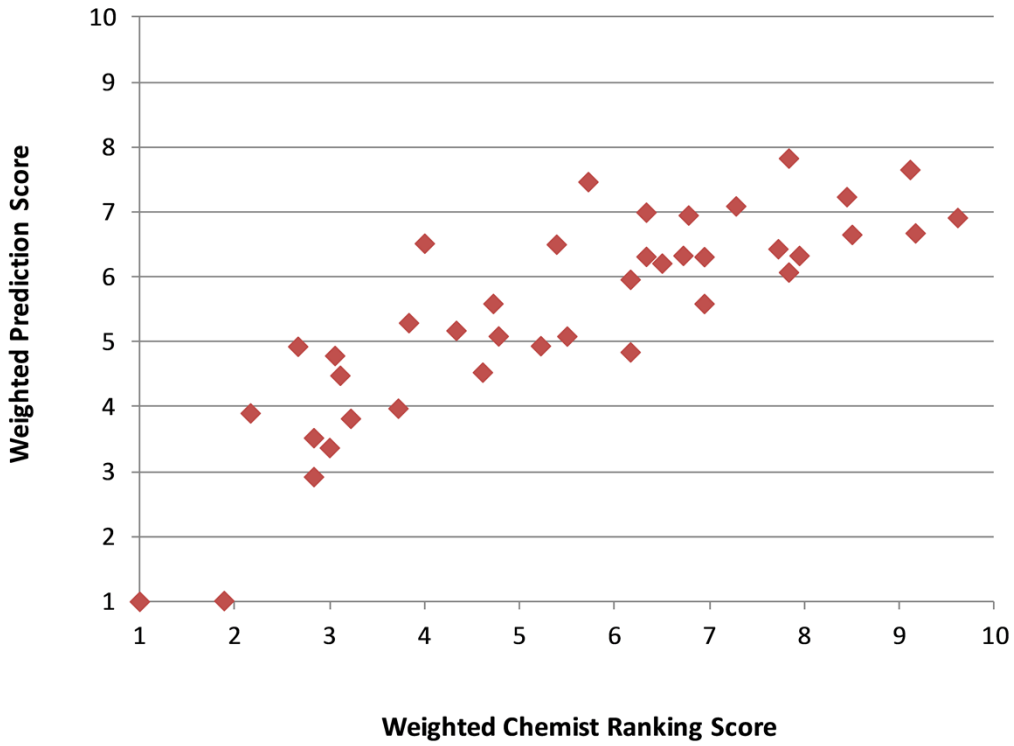


Figure S8. Weighted complexity prediction versus weighted chemist ranking score

	Randic	Steps	Ideality	Chiral_made	HAA	predicted complexity score	
Topiramate	10.99264	3	0.67	0	0	7.83	
Sidenafil	15.75559	8	0.89	0	7	6.42	
Atazanavir	25.40878	8.5	0.6	0	1	6.27	
BMS-2	23.16366	11	0.56	1	4	5.16	
BMS-1	18.35124	16	0.56	3	3	4.04	
BILN2061	27.91184	19.5	0.54	3	6	2.88	
EpoA	18.26984	27	0.47	5	2	2.38	Danishefsky
Strychnine	13.72358	8.5	0.60	6	1	3.75	Vanderwal
Strychnine	13.72358	13	0.69	6	1	3.43	MacMillan
Strychnine	13.72358	25	0.56	6	1	2.45	Overman
Strychnine	13.72358	30	0.53	6	1	2.14	Woodward
Discodermolide	23.87016	30	0.59	10	0	1.25	Novartis-Smith-Paterson
Taxol	31.71272	37	0.48	11	0	1.06	Nicolaou
Halaven	31.3614	56	0.38	13	0	1.00	Eisai
Welwit C isonitrile	13.3244	21	0.33	4	1	3.42	Rawal
Welwit C isonitrile	13.3244	23	0.39	4	1	3.30	Garg
Welwit A isonitrile	12.45351	23	0.43	4	1	3.38	Wood
Welwit A isonitrile	12.45351	9	0.78	4	1	4.98	Baran

Table S8. Input parameters for selected molecules and weighted predicted complexity score

---

[1] a) M. Grabner, K. Varmuza, M. Dehmer, *Source Code for Biology and Medicine* **2012**, *7*, 12; b) L. A. J. Mueller, M. Schutte, K. G. Kugler, M. Dehmer, *QuACN: Quantitative Analyze of Complex Networks* **2012**.
[2] T. Lumley, R package leaps **2013** from CRAN R project.

[3] a) WinBUGS: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml; b) OpenBUGS: http://www.openbugs.net/w/FrontPage.
[4] R Core Team **2014**. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/
[5] A. D. Martin, K. M. Quinn, J. H. Park, R package MCMCpack **2013** from CRAN R project.

[6] K. Kruschke, *Doing Bayesian Data Analysis.* **2011**, Elsevier.
[7] a) P. Ertl, A. Schuffenhauer, *J. Cheminformatics* **2009**, *1,* 8; b) K. Boda, T. Seidel, J. Gasteiger, *J. Comput. Aided. Mol. Des.* **2007**, *21,* 311.