ACS Combinatorial Science

RESEARCH ARTICLE

pubs.acs.org/acscombsci

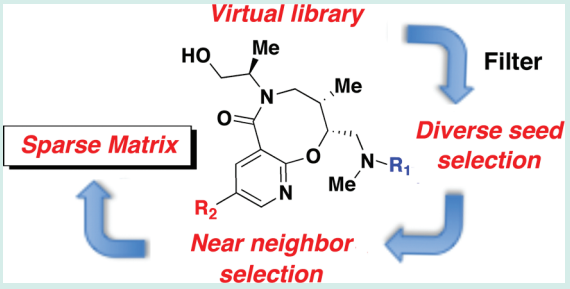
Application of a Sparse Matrix Design Strategy to the Synthesis of DOS Libraries

Lakshmi B. Akella and Lisa A. Marcaurelle[†,*]

Chemical Biology Platform, The Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02142, United States

Supporting Information

ABSTRACT: We have implemented an interactive and practical sparse matrix design strategy for the synthesis of DOS libraries, which facilitates the selection of diverse library members within a user-defined range of physicochemical properties while still maintaining synthetic efficiency. The utility of this approach is illustrated with the synthesis of an 8000-membered library of stereochemically diverse medium-sized rings accessible via a build/couple/pair DOS strategy. Diverse library members were selected from a virtual library by applying the maximum dissimilarity method, while the selection of similar analogs around each diverse product was ensured by picking near neighbors algorithmically based on fingerprint comparison. Adjustable filters on compound properties, which can be tailored to suit the needs of the target biology, facilitated subset selection from the synthetically accessible compounds.



KEYWORDS: library design, diversity-oriented synthesis, physicochemical properties, diversity-ranking, maximum dissimilarity, sparse matrix

INTRODUCTION

Designing libraries with properties suitable for use in biological screens and downstream discovery is a critical step in the synthesis of any compound collection.[1] It has been noted that compounds resulting from diversity-oriented synthesis (DOS) often violate Lipinski's Rule of 5[2] with high molecular weight and predicted low solubility.[1,3] The outcome of any library synthesis, however, is a product of the design. A primary goal of DOS is the synthesis of skeletally diverse small molecules of increased structural complexity (e.g., high sp^3 content, multiple stereogenic centers) that can be accessed in relatively few synthetic steps.[4] If effort is taken up front to control the physicochemical properties of DOS library members, these structural features can be achieved while still producing compounds with favorable physicochemical properties.

When designing a small-molecule library for high-throughput screening, chemists are faced with the challenge of selecting which compounds to synthesize. There exists abundant literature on various computational library design methods to choose an optimal subset for synthesis (or screening) from large chemical spaces.[5] The two main approaches are reagent- and product-based design.[6,7] We have combined both of these design strategies for the synthesis of DOS libraries, which we illustrate here with an 8000-membered library of stereochemically diverse medium-sized rings. The DOS scaffold that was selected as a starting point for library design is shown in Figure 1, along with a set of structurally related scaffolds.[8] The latter will be used to compare differences in design strategies as they influence the physicochemical property profile of library members (vide infra). The process employed for the selection of the scaffold itself involved the use of various methods commonly employed for assessing diversity (e.g., Principal Moments of Inertia, Multi-fusion Similarity, Principal Component Analysis).[9] The synthesis of the S_NAr-Pyr scaffold and its corresponding stereoisomers is the subject of a separate communication.[10]

Three approaches have been utilized for the design of compound libraries: (1) a *full matrix* design where every reagent at R_1 is combined with every reagent at R_2 thus forming the maximum number of products, (2) a *sparse matrix* design strategy[5b,11] where a subset of products are selected for synthesis and not all reagent combinations are selected, and (3) a *cherry pick* strategy[5c] where a subset of products (diverse or similar) is selected for synthesis. Although we have utilized a full matrix approach for previous DOS efforts[8] a sparse matrix design strategy appealed to us for the purpose of controlling the physicochemical properties of the library members as well as maximizing coverage of chemical space. While diverse chemical space with suitable physicochemical properties can be achieved through a cherry pick approach (using a combination of property-based filtering and diversity-ranking) we generally reserve

Received: February 1, 2011
Revised: April 6, 2011
Published: April 28, 2011

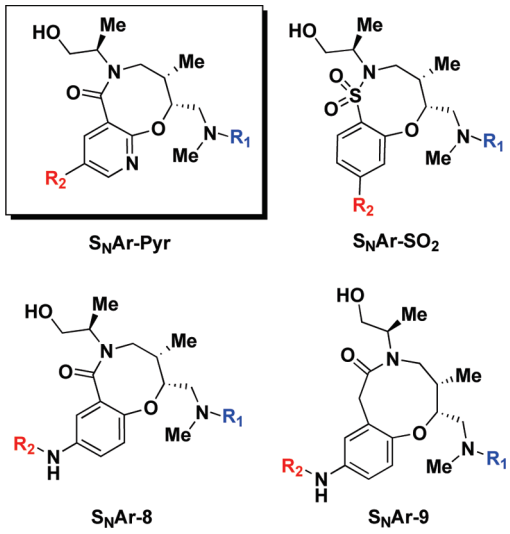


Figure 1. S_NAr-based DOS library scaffolds (R_1 = diversity site 1, R_2 = diversity site 2).

this method for small compound libraries (<100 compounds). We envisioned that a sparse matrix approach would allow for the selection of "near neighbors" around each diverse molecule thus facilitating access to built-in structural analogs in contrast to a diversity-ranking approach. In this regard a sparse matrix design achieves a balance between a full matrix design and cherry picking.

The design process we have implemented involves the following: (1) creation of master lists for the various reagent classes, (2) library enumeration based on defined production pathways, and (3) compound selection using the sparse matrix approach. The design is carried out on a single stereoisomer and applied to all the other stereoisomers[12] thereby maintaining the ability to generate stereo/structure−activity relationships upon biological testing.[8] Details of the design workflow are outlined below.[13]

RESULTS AND DISCUSSION

Master Reagent Lists. Before a virtual library could be created for product-based selection, a list of suitable building blocks for each reagent class was required. As selecting reagents intuitively from large databases is an arduous task, several software tools and systems have been developed to aid in the filtering process.[11,14] At our end we have implemented a reagent selection process that takes into consideration diversity (with respect to structure and properties), synthetic feasibility, availability, and price. As shown in Figure 2 reagents for various reagent classes were retrieved from Available Chemicals Directory (ACD) by using functional group queries. Reagent classes included in our search were sulfonyl chlorides, isocyanates, aldehydes, acids, acid chlorides, alkynes, boronic acids, and amines. The resulting reagents were filtered by molecular weight (≤200) and number of rotatable bonds (≤5) and then exported as structure definition files (sd files). The first step of the filtering process involved the stripping of salts and removal of duplicate molecules. General exclusion filters that were applicable to all reagent classes were created in the form of Daylight SMARTS queries. These filters include isotopes, inorganic elements, excessive number of halogens, charged species, peroxides, thiols, Michael acceptors, etc.[15]

Reagent class specific filters were then applied to the reagent lists. For example, carboxylic acid specific exclusion filters

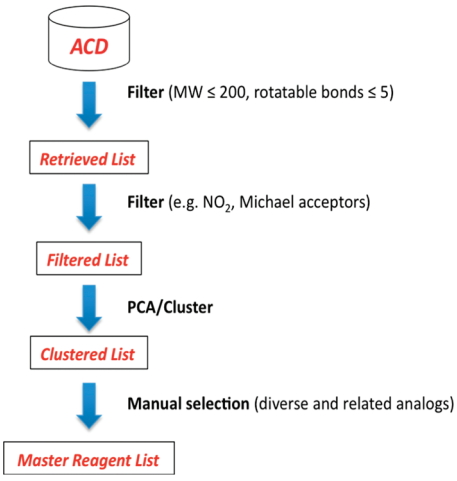


Figure 2. Workflow for the creation of master lists for various reagent classes.

Table 1. Reagent List Generation

reagent class	total	filtered	no. clusters	no. selected	no. S_NAr-Pyr[a]
sulfonyl chlorides	171	153	30	23	11
isocyanates	517	438	80	20	15
aldehydes	5953	4307	120	46	22
acids	17 935	8551	200	26	24
acid chlorides	1008	780	60	30	N/A
alkynes	1752	1270	120	23	23
boronic acids	1539	846	60	23	20
amines	39 453	21 504	200	60	N/A

[a] Not all reagents on the master list were used for library enumeration. (See Supporting Information).

included primary or secondary amines, formyl, nitro, nitroso, carboxyl count > 2, isocyanate, imino, allene, epoxide, anhydride, etc. The successive filtering/eliminations resulted in a manageable and significantly reduced list for each reagent classes of interest. Various structural and physicochemical properties such as molecular weight, ALogP, topological polar surface area (TPSA), number of acceptors, number of donors, number of rotatable bonds, number of rings, number of ring assemblies and ring size were calculated. Principal component analysis (PCA) was performed on the properties,[16,17] followed by clustering on principal components using the maximum dissimilarity method[18] for selecting cluster centers. The number of clusters is predefined (Table 1) depending on the reagent class size.

Each reagent class was created as a project in Instant JChem (ChemAxon) and cluster centers were automatically marked to facilitate reagent selection. Chemists visually inspected the clusters and selected reagents manually (not always the cluster center) mindful of reactivity, synthetic feasibility, price, and availability.[19,20] Many clusters resulted in no selections due to lack of availability or medicinal chemistry considerations. When a particular structure was selected from a given cluster a small number of closely related analogs were also chosen from the same cluster to ensure SAR. Final selections consisted of 20−50 reagents per class. We periodically update the master reagent lists based on the

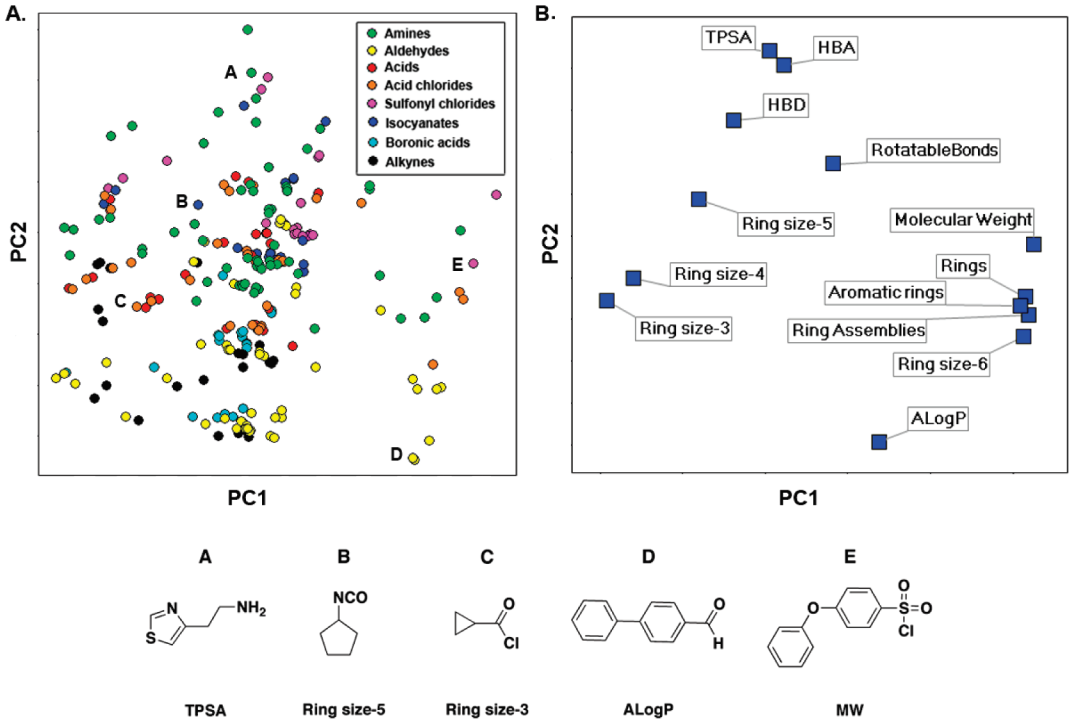
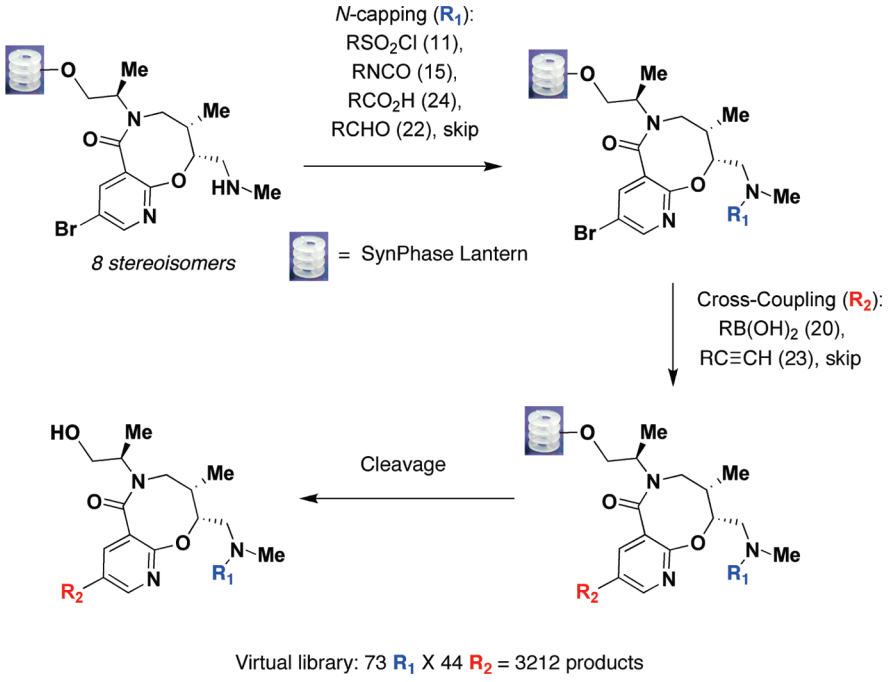


Figure 3. (a) Principal component analysis (PCA) for reagent master lists based on fragments properties. (b) Loading plot displaying properties used in the analysis. Representative reagents are provided (**A**-**E**) to illustrate which properties influence their location on the PCA plot. (TPSA = topological polar surface area, HBD = number of hydrogen bond donors, HBA = number of hydrogen bond acceptors, Aromatic rings = number of aromatic rings, Rings = number of rings, Ring assembly = connectivity of rings).

Scheme 1. Solid-Phase Synthesis Plan for S_NAr-Pyr Library



synthetic outcomes or commercial availability. Our most current master reagent lists are provided in the Supporting Information.

The fragment property space for the master list of selected reagents can be visualized using a PCA plot (Figure 3A). Reagents that are close to each other on the plot are similar. The loadings plot (Figure 3B) shows the relationship between properties, which dictate the location of the fragment on the PCA plot. For example, properties that are negatively correlated such

dx.doi.org/10.1021/co200020j |*ACS Comb. Sci.* 2011, 13, 357–364

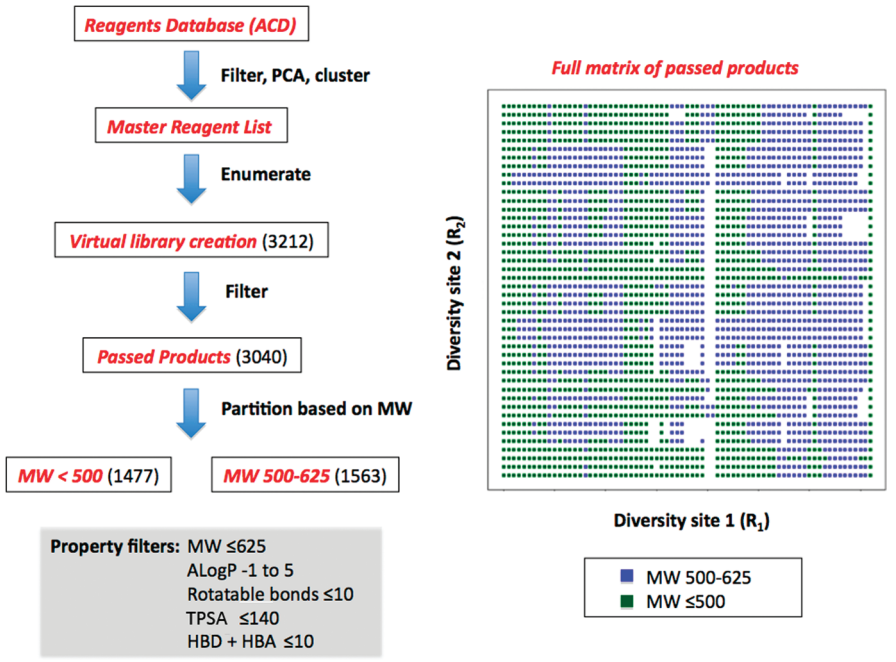


Figure 4. Workflow for in silico library enumeration and product filtering process.

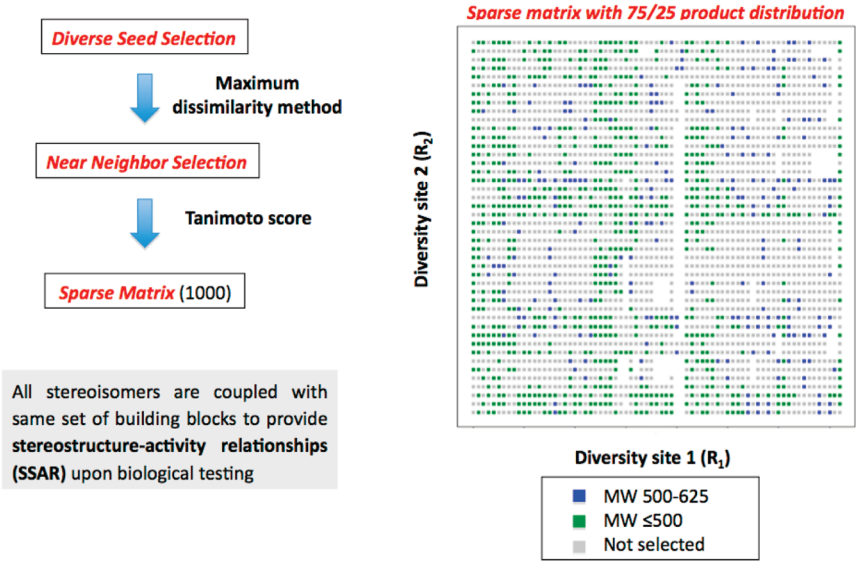


Figure 5. Workflow for in silico compound selection.

as TPSA and ALogP (which influence polarity and hydrophobicity respectively) appear on opposite sides of the plot. When in need of back-up reagents (because of synthetic feasibility or unavailability), the selection of close analogs is facilitated by revisiting the clustered reagent lists and PCA plot. We anticipate these larger yet selected spaces to be useful upon screening when hits are identified and additional SAR expansion is required.

Library Enumeration and Product Filtering. With the master lists of reagents in hand, a virtual library was constructed for the S_NAr-Pyr scaffold where every reagent at R_1 was used in all combinations with reagents at R_2 thereby resulting in a full combinatorial matrix. All synthetically accessible production pathways were enumerated (including "skips" at R_1 or R_2). The synthetic sequence for the S_NAr-Pyr library is shown in Scheme 1. The reagent classes used at R_1 included sulfonyl chlorides, isocyanates, acids and aldehydes while reagents used at R_2 included boronic acids and alkynes for Suzuki and Sonogashira reactions respectively. Enumeration was reaction based; a full list of SMIRKS used for enumerating the S_NAr-Pyr library are provided as Supporting Information. The total number of enumerated products for the S_NAr-Pyr library is 3212 compounds (72 reagents (+ 1 skip) at R_1 × 43 (+ 1 skip) reagents at R_2). (Reagents at R^1 containing aryl chlorides were removed because of incompatibility with the cross coupling step.) The in silico library enumeration and product filtering process is depicted in Figure 4.

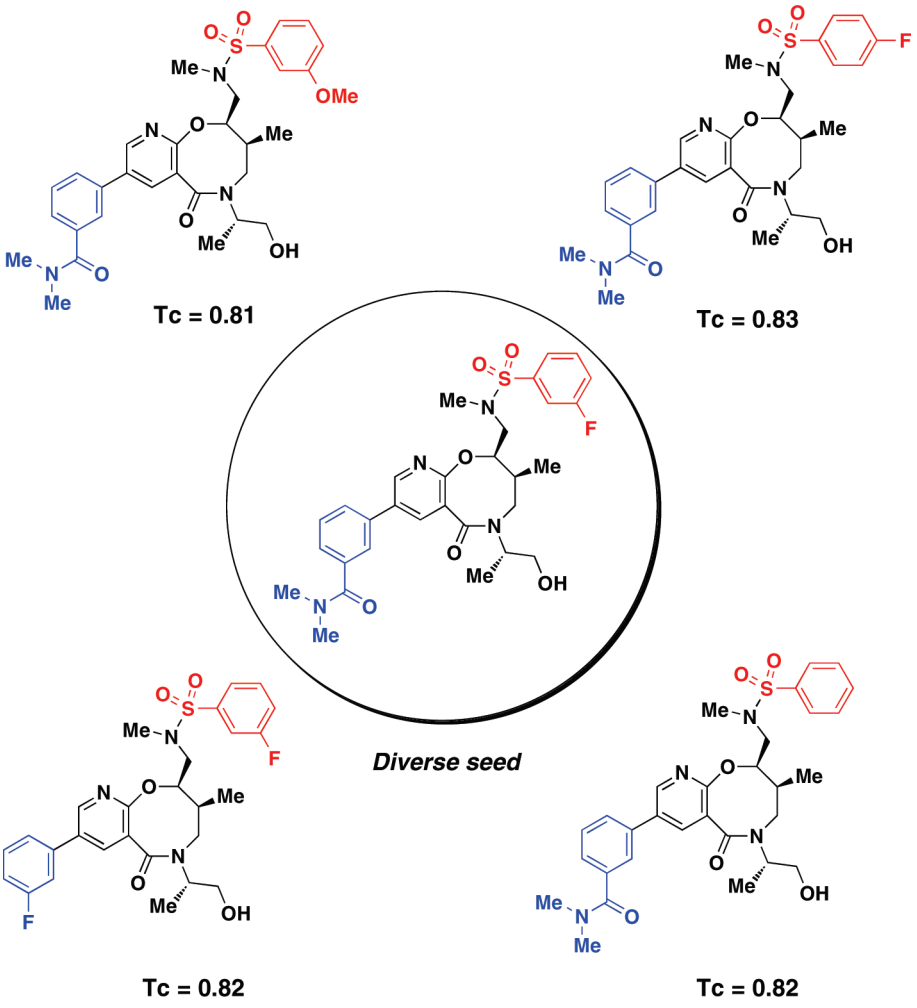


Figure 6. Representative diverse seed for S_NAr Pyr library and selected four near neighbors (Tc = Tanimoto coefficient).

Next, molecular properties that affect solubility, permeability, and bioavailability were calculated for each product.[1,21] The properties and the threshold limits applied on the products are molecular weight (≤625), AlogP (−1 to 5), number of H-bond acceptors plus donors (≤10), number of rotatable bonds (≤10), and topological polar surface area (≤140). Structures that violated any single property were eliminated. We implemented a "75/25" rule where the data set was partitioned into two data streams based on molecular weight: less than 500 and greater than 500. This rule was applied to favor products with molecular weight less than 500, while still allowing for a small percentage of "Lipinski violators" to be formed. The 75/25 rule is applied after reviewing the enumerated chemical space. If the enumerated chemical space largely occupies molecular weight <500 then the 75/25 rule is not applied. In the case of the S_NAr-Pyr library the space is roughly equally distributed with respect to molecular weight partitioning.

Sparse Matrix Design. After filtering based on properties, a subset of products is chosen from the virtual library based on chemical similarity principles.[22] The chemical similarity principle assumes that structurally similar compounds should have similar biological activity.[23] As shown in Figure 5, diverse molecules were selected from each partitioned data set based on the maximum dissimilarity method.[14] The number of diverse molecules is user defined and is dependent on the library size. In this instance a 1000-membered library was desired. For every diverse molecule at most four near neighbors were selected algorithmically based on pairwise fingerprint similarity of the structures with a similarity

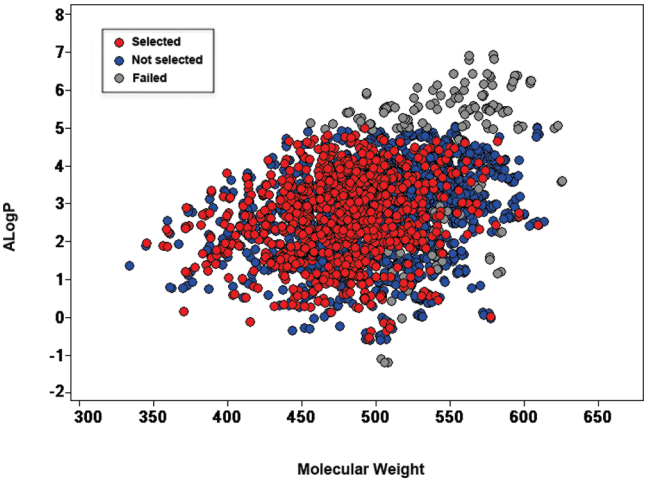


Figure 7. Scatter plot of molecular weight vs ALogP for the S_NAr-Pyr library, selected (1000), not selected (2040) and failed (172).

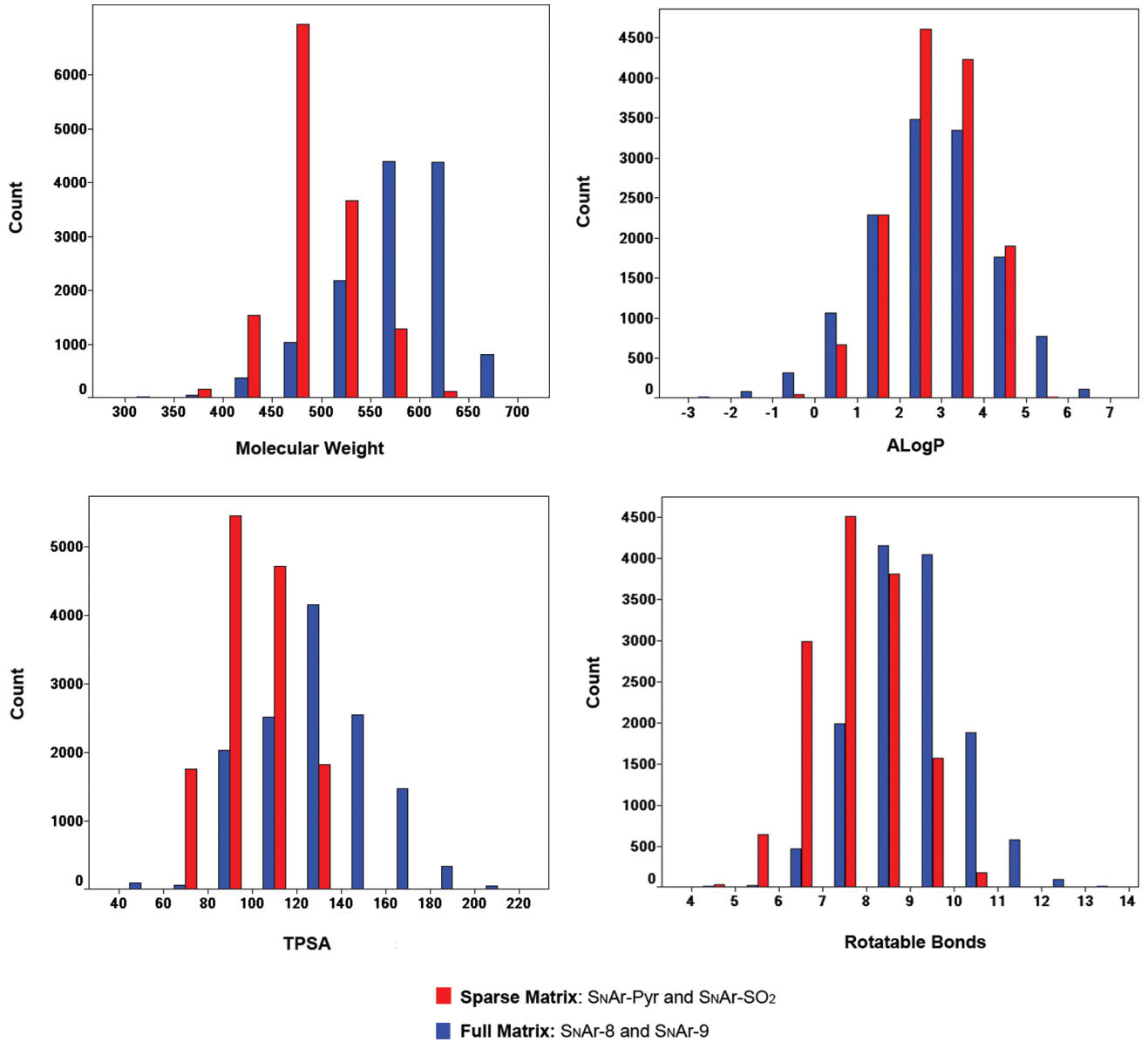


Figure 8. Property distribution for S_NAr-based DOS libraries: Full (blue) vs sparse (red) matrix design.

threshold (Tanimoto coefficient, Tc) of 0.8 (Figure 6).[24,25] A molecule already considered as a neighbor is dropped from future selection. The number of reagents selected for the S_NAr-Pyr library production includes 24 acids, 22 aldehydes, 15 isocyanates, and 11 sulfonyl chlorides, 20 boronic acids and 23 alkynes.[26] For this particular library, relatively few reagents were dropped given the proportion of selected to enumerated space (1:3). For smaller libraries (or those with more than two diversity sites) a larger number of reagents tend to be dropped.

To achieve synthetic efficiency during library production we typically set a minimum threshold for the number of products formed per reagent. This is done to prohibit the selection of reagents that form only a small number of products. The threshold is user defined and can vary by enumerated library size. For the S_NAr-Pyr library, a minimum count of 5 products per reagent was applied. On review of the outcome of the design, excessive use of any one reagent is curtailed by applying a limiting filter and the design is repeated accordingly. If a problematic reagent is identified during feasibility studies that reagent can be dropped along with the associated products. In such situations new products are selected from the remaining chemical space that are dissimilar to already selected products.

Property Analysis. Following the implementation of the sparse matrix design we analyzed the selected product space with respect to molecular weight and ALogP. As shown in Figure 7 compounds selected for synthesis are uniformly distributed across the virtual chemical space with a greater number of compounds occupying the region MW <500. Also shown in the plot are compounds that passed the property criteria but were not selected and compounds that failed any single property filter (MW, ALogP, TPSA, rotatable bonds, etc).

Lastly we compared the property distribution of compounds resulting from a sparse vs full matrix design strategy in the context of a set of structurally related DOS scaffolds (structures shown in Figure 1). Similar to the S_NAr-Pyr library a sparse matrix design strategy was applied to the S_NAr-SO_2 scaffold. Meanwhile a full

Table 2. Property Analysis for S_NAr-Based DOS Libraries: Full versus Sparse Matrix Design

property	full matrix (n = 13 270)	sparse matrix (n = 13 735)
MW	576	494
ALogP	2.8	2.8
TPSA	129	100
rotatable bonds	8.5	7.2
HBA	6.6	6.0
HBD	2.6	1.3

matrix strategy was employed for the S_NAr-8 and S_NAr-9 scaffolds. (In the latter case, no property filters were applied.) The combined property analysis is shown in Figure 8. A mere visual inspection shows a clear shift in distribution in the desired direction for all properties, especially molecular weight and polar surface area. Mean values calculated for each of the descriptors at the library level also reflect the same (see Table 2). Notably, for this particular set of scaffolds the mean ALogP, HBD and HBA values were deemed acceptable even without the sparse matrix design.

CONCLUSION

In summary, we have implemented a reagent- and product-based sparse matrix design strategy that is both interactive and practical, involving full participation of the chemists. The key features of the compound selection are desirable physicochemical properties, diversity and built-in structural analogs and synthetic efficiency. We expect our design-synthesis-screening cycle to inform future library design and suggest refinements to our approach. As the product filters can be adjusted at the outset of the design, the property profile of the library can be tailored to meet the needs of the therapeutic area of interest (e.g., CNS, antibacterial).[27]

ASSOCIATED CONTENT

Supporting Information. Methods and software tools used for data processing are provided as well as master reagents lists and SMIRKS used for enumeration.This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: lisa_marcaurelle@h3biomedicine.com.

Present Addresses

†Current address: H3 Biomedicine Inc., 300 Technology Square, Cambridge, MA 02139.

ACKNOWLEDGMENT

The authors would like to acknowledge the members of the Chemical Biology Platform who influenced the overall design strategy described here, including Dr. Michael Foley, Dr. Benito Munoz, Dr. Lawrence MacPherson, Dr. Jeremy Duvall and Phillip Montgomery. This work was funded in part by the NIGMS-sponsored Center of Excellence in Chemical Methodology and Library Development (Broad Institute CMLD; P50 GM069721), as well as the NIH Genomics Based Drug Discovery U54 grants Discovery Pipeline RL1CA133834 (administratively linked to NIH grants RL1HG004671, RL1GM084437, and UL1RR024924).

REFERENCES

(1) Shelat, A. A.; Guy, K. The interdependence between screening methods and screening libraries. *Curr. Opin. Chem. Biol.* **2007**, *11*, 244–251.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.

(3) (a) Cooper, T. W. J.; Campbell, I. B.; Macdonald, S. J. F. Factors determining the selection of organic reactions by medicinal chemists and the use of these reactions in arrays (small focused libraries). *Angew. Chem., Int. Ed.* **2010**, *49*, 8082–8091. (b) Hajduk, P. J.; Galloway, W. R. J. D.; Spring, D. R. Drug discovery: A question of library design. *Nature* **2011**, *470*, 42–43.

(4) (a) Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287* (5460), 1964–1969. (b) Burke, M. D.; Schreiber, S. L. A planning strategy for diversity-oriented synthesis. *Angew. Chem., Int. Ed.* **2004**, *43*, 46–58. (c) Neilson, T. E.; Schreiber, S. L. Towards the optimal screening collection: A synthesis strategy. *Angew. Chem., Int. Ed.* **2007**, *74*, 48–56.

(5) (a) Hobbs, D. W.; Guo, T. Library design concepts and implementation strategies. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 1−49. (b) Weber, L. Current status of virtual combinatorial library design. *QSAR Comb. Sci.* **2005**, *24*, 809–823.(c) Brown, R. D.; Hassan, M.; Waldman, M. Tools for designing diverse, druglike, cost-effective combinatorial libraries. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 301−335. (d) Blake, J. F. Integrating cheminformatic analysis in combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2004**, *8*, 407–411.

(6) Jamois, E. A. Reagent-based and product-based computational approaches in library design. *Curr. Opin. Chem. Biol.* **2003**, *7*, 326–330.

(7) Reagent-based design considers the properties of the reagent while product-based design considers the properties of the whole molecule.

(8) Marcaurelle, L. A.; Comer, E.; Dandapani, S.; Duvall, J. R.; Gerard, B.; Kesavan, S.; Lee, M. D., IV; Liu, H.; Lowe, J. T.; Marie, J-.C.; Mulrooney, C. A.; Pandya, B. A.; Rowley, A.; Ryba, T. D.; Suh, B-. C.; Wei, J.; Young, D. W.; Akella, L. B.; Ross, N. T.; Zhang, Y-. L.; Fass, D. M.; Reis, S. A.; Zhao, W-. Z.; Haggarty, S. J.; Palmer, M.; Foley, M. A. An aldol-based build/couple/pair strategy for the synthesis of medium- and large-sized rings: Discovery of macrocyclic histone deacetylase inhibitors. *J. Am. Chem. Soc.* **2010**, *132*, 16962–16976.

(9) The S_NAr-Pyr scaffold has been used to illustrate various methods for diversity analysis that we typically employ for scaffold selection, see: Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325–330.

(10) Gerard, B.; Duvall, J. R.; Lowe, J. T.; Murillo, T.; Wei, J.; Akella, L. B.; Marcaurelle, L. A. Synthesis of a stereochemically diverse library of medium-sized lactams and sultams via S_NAr cycloetherification. *ACS Comb. Sci.* **2011**10.1021/co2000218.

(11) Clark, R. D.; Kar, J.; Akella, L.; Soltanshahi, F. OptDesign: Extending optimizable *k*-dissimilarity selection to combinatorial library design. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 829–836.

(12) The same approach is applied to reagents, where the stereochemistry of the reagent is ignored during enumeration but applied during production.

(13) Methods and software tools used for data processing are provided in the Supporting Information.

(14) (a) Yasri, A.; Berthelot, D.; Gijsen, H.; Thielemans, T.; Marichal, P.; Engels, M.; Hoflack, J. REALISIS: A medicinal chemistry-oriented reagent selection, library design, and profiling platform. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2199–2206. (b) Valerie, J.; Gillet,

J. V.; Willett, P.; Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731–740. (c) Truchon, J. F.; Bayly, I. C. GLARE: A new approach for filtering large reagent lists in combinatorial library design using product properties. *J. Chem. Inf. Model.* **2006**, *46*, 1536–1548. (d) Rhodes, N.; Willett, P.; Dunbar, J. B., Jr.; Humblet, C. Bit-strings methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 210–214.

(15) A full list of general filters can be found in the Supporting Information.

(16) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002.

(17) Principal component analysis is a data reduction technique where a number of correlated variables are transformed into a smaller number of uncorrelated variables. Each transformation is a principal component and related to the other components orthogonally, the first component explaining the maximum variance in the data and so on. Each descriptor was centered (mean was subtracted from each property value) and scaled (each property value was divided by the variance) prior to the analysis.

(18) (a) Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41*, 478–488. (b) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64–74.

(19) See Supporting Information for an example.

(20) The price and availability data is inherently noisy because of the variability in quantity (i.e., price per gram) and annotations such as POA (price on asking), making it difficult to automate filtering by these attributes.

(21) Veber, D .F.; Johnson, S. R.; Cheng, H-. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(22) Mosley, T. R.; Culberson, J. C.; Kraker, B.; Feuston, P. B.; Sheridan, P. B.; Conway, F. J.; Forbes, K. J.; Chakravorty, J. S.; Kearsley, K. S. Reagent selector: Using synthon analysis to visualize reagent properties and assist in combinatorial library design. *J. Chem. Inf. Model.* **2005**, *45*, 1439–1446.

(23) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual compound libraries: A new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010–1023.

(24) The molecular similarity matrix between each molecule and all the other molecules in the dataset is computed based on extended connectivity fingerprints (ECFP_4), see: Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(25) The similarity between each pair is recorded as Tanimoto coefficient (Tc), see: Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.

(26) A list of final building blocks selected for library production are provided in Supporting Information.

(27) (a) Hitchcock, S. A.; Pennington, L. D. structure−brain exposure relationships. *J. Med. Chem.* **2006**, *49*, 7559–7583. (b) O'Shea, R.; Moser, H. E. Physicochemical properties of antibacterial compounds: Implications for drug discovery. *J. Med. Chem.* **2008**, *51*, 2871–2878.