ACS Combinatorial Science

Research Article

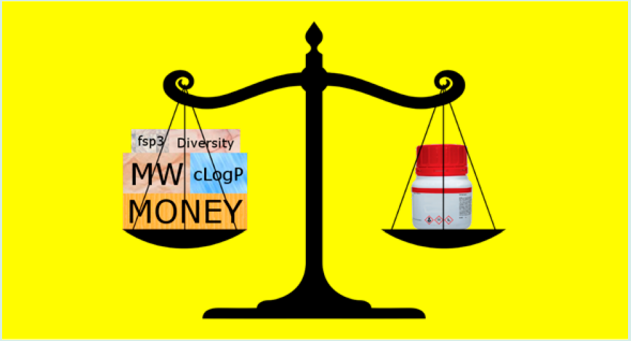
pubs.acs.org/acscombsci

Price-Focused Analysis of Commercially Available Building Blocks for Combinatorial Library Synthesis

Tuomo Kalliokoski*

Lead Discovery Center GmbH, Otto-Hahn-Straße 15, 44227 Dortmund, Germany

ABSTRACT: Combinatorial libraries are synthesized by combining smaller reagents (building blocks), the price of which is an important component of the total costs associated with the synthetic exercise. A significant portion of commercially available reagents are too expensive for large scale work. In this study, 13 commonly used reagent classes in combinatorial library synthesis (primary and secondary amines, carboxylic acids, alcohols, ketones, aldehydes, boronic acids, acyl halides, sulfonyl chlorides, isocyanates, isothiocyanates, azides and chloroformates) were analyzed with respect to the cost, physicochemical properties (molecular weight and calculated logP), chemical diversity, and 3D-likeness using a large data set. The results define the chemical space accessible under a constraint of limited financial resources.



KEYWORDS: *building blocks, price, cost, diversity, combinatorial library, library design*

INTRODUCTION

The synthesis of candidate compounds by combining small sets of reagents (building blocks)[1] has become a standard tool in drug discovery.[2] Molecular weight and calculated logP are often used to guide computational combinatorial library design, with some justification.[3,4] The additive or iterative use of such descriptors for weeding out unsuitable building blocks has been demonstrated in rapid methods such as GLARE in the library design process.[5] In addition to these two simple criteria, substructural filters are applied to rid libraries of unwanted functional groups that may cause problems in either stability or later in biological assays.[6,7] More abstract features of molecules can also be considered in the design phase as well, such as "3D-likeness" to known effective motifs,[8,9] with an eye toward avoiding flat structures that are prone to nonspecific binding.[10] However, not often included in presynthesis algorithms are the two practical criteria of availability in a reasonable time frame and cost.

For example, the generation of a 1000-member compound library around a scaffold that has three different diversification points might start with four initial building blocks, which are decorated with ten variants of the second block and 25 variants for the third ($4 \times 10 \times 25 = 1000$). The components used earlier in the synthesis are required in larger amounts; if expensive building blocks are used, the price of the library shoots up very quickly. Therefore, information on the relation between the price and the properties of a building block becomes significant when the aim is to produce large combinatorial libraries. One such undertaking is part of the European Lead Factory (ELF)-project[11] where the author is currently employed. In this project, there is work in progress of synthesizing 200 000 novel and medicinal chemistry-friendly compounds for high-throughput screening. The production of such a large number of compounds requires efficient use of limited resources so that the relevant chemical space is explored as widely as possible.

The aim of this study was to investigate how much of the commercially available building block chemical space is reachable when one takes the price of the reagents into account and when it is useful to consider also the more expensive reagents to increase the diversity of a library or to improve the physicochemical properties. It is surprisingly difficult to find such studies in the published literature, although they are a routine part of internal assessment in the commercial drug discovery sector. One reason for the absence of such studies could be the fact that price and availability information is rather laborious to collect and to keep up to date. In practice, such information is only available via proprietary databases like Accelrys Available Chemicals Directory (ACD)[12] or eMolecules Plus.[13] The interest in the scientific literature on building blocks seem to focus on Synthetic Accessibility (SA) of the resulting compounds and from the scientific literature only anecdotal observations between price and chemical structures can be found, such as *"We were not able to get any reasonable correlation between normalized catalogue price and various structural descriptors for a large set of reagents"*.[14] During the preparation of this article, a group of researchers from AstraZeneca published a study on the preference of using para-substituted

Received: April 27, 2015
Revised: June 22, 2015
Published: September 15, 2015

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

phenyl rings instead of ortho- and meta-substituted ones. Their study also looked at the cost of the building blocks as a factor for the observed preference and concluded that cost was not a likely reason for the bias.[15]

This study provides the first published comprehensive analysis of commercially available building blocks, with special focus on the price per gram-cost of the reagents. Thirteen different commonly used reagent classes were analyzed: primary and secondary amines, carboxylic acids, alcohols, ketones, aldehydes, boronic acids, acyl halides, sulfonyl chlorides, isocyanates,isothiocyanates, azides, and chloroformates. Obviously, there are building blocks available that do not fall into these classes, but these are the reagent classes that are the most often used ones in the ELF-project and thus should be typical for any large scale library synthesis undertaking. The physicochemical properties of molecular weight and calculated logP, chemical similarity, and 3D-likeness of the compounds are investigated and discussed.

RESULTS AND DISCUSSION

From Table 1, it is clear that the price distributions of building blocks are strongly skewed and therefore, the median price should be used instead of the mean price when discussing the general prices of building blocks. Also, a large portion of commercially available primary amines and alcohols are prohibitively expensive. One needs to have price and availability information readily available when designing libraries as otherwise it is very easy to pick building blocks that are too costly.

Amines are the largest group of commercially available building blocks, and so scaffolds that can use primary and/or secondary amine reagents in their diversification points are attractive from the diversity point of view. On other hand, there are only few chloroformates, azides, isothiocyanates, and isocyanates available. These diversifications therefore offer limited potential for the chemical space exploration.

The medians of price per gram for several reagent classes after the price-cutoff of 1000 USD per gram are still rather high from a practical point of view. Therefore, several lower price cutoffs instead of the median, which better reflect the reality of combinatorial library synthesis, were applied to the analysis: 25, 50, 100, and 150 USD per gram (Table 2). Azides are expensive building blocks as less than one-third are available for 150 USD per gram, whereas chloroformates are cheap as most of them can be bought less than 25 USD per gram.

The relationships between building block prices and properties are not simple, as the quote from the study by Ertl and Schuffenhauer in the Introduction illustrated. Similarly for the data set used in this study, there was no correlation observed in any reagent class between price and molecular weight or calculated logP when looking at the data set of maximum 1000 USD per gram (Pearson correlation coefficient ranged between −0.20 and 0.35, calculated with RDKit and R, data not shown). However, some useful observations could be made by binning the building blocks using price, molecular weight and calculated logP (Table 3). This table can be used to quickly check if there are any possibilities in certain molecular weight/logP range for certain reactions without resorting to performing extensive calculations and to see if the limited chemical space can be expanded by consideration of more expensive reagents. As an example on how Table 3 could be used: it would be difficult to find secondary amines that have calculated logP between 2.0 and 4.0 and that weigh between 100 and 150 Da, even if one would consider more expensive

Table 1. Number of Building Blocks Used in This Study[a]

class	number of BB	excluded BB	mean price ($/g)	median price ($/g)
primary amines	79142	27020	333	231
secondary amines	51889	9479	254	227
carboxylic acids	44760	5606	223	174
alcohols	28021	18374	239	200
ketones	21600	3835	211	139
aldehydes	11930	1307	227	161
boronic acids	4293	525	220	129
acyl halides	2180	116	172	104
sulfonyl chlorides	1662	57	212	194
isocyanates	753	31	152	84
isothiocyanates	713	35	105	58
azides	652	116	245	213
chloroformates	77	2	68	8

[a]The rows are sorted by the number of building blocks. Building blocks that cost over 1000 USD per gram were excluded (the numbers are reported in the excluded BB-column).

Table 2. Building Block Availability at Lower Price Cutoffs[a]

	$25/g		$50/g		$100/g		$150/g	
class	count (%)	median price ($/g)	count	median price ($/g)	count	median price ($/g)	count	median price ($/g)
primary amines	10021 (13%)	3	13346 (17%)	7	19084 (24%)	22	27028 (34%)	52
secondary amines	6796 (13%)	6	9239 (18%)	15	12818 (25%)	21	17611 (34%)	46
carboxylic acids	7983 (18%)	3	10389 (23%)	6	14907 (33%)	20	20531 (46%)	48
alcohols	4565 (16%)	3	5786 (21%)	6	7888 (28%)	16	11925 (43%)	54
ketones	4043 (19%)	4	5155 (24%)	7	7824 (36%)	23	11328 (52%)	64
aldehydes	2436 (20%)	3	3242 (27%)	7	4333 (36%)	18	5804 (49%)	39
boronic acids	1107 (26%)	3	1444(34%)	7	1859 (43%)	16	2310 (54%)	29
acyl halides	670 (31%)	3	818 (38%)	6	1055(48%)	14	1420 (65%)	29
sulfonyl chlorides	286 (17%)	3	354 (21%)	5	490 (29%)	15	662 (40%)	38
isocyanates	207(27%)	6	308 (41%)	14	406 (54%)	24	488 (65%)	32
isothiocyanates	282 (40%)	4	350 (49%)	7	439 (62%)	13	552 (77%)	24
azides	45 (7%)	7	60 (9%)	9	95 (15%)	29	182 (28%)	97
chloroformates	48 (62%)	1	56 (73%)	3	60 (78%)	5	62 (81%)	5

[a]Count is the number of the building blocks, and median is the median price of building blocks at the cutoff. The percentage is the percentage of the building blocks at this cutoff from the set of building blocks available for 1000 USD/gram.

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

Table 3. Numbers of Building Blocks Clustered by Molecular Weight (W) and Calculated logP (P)[a]

(a) primary amines

	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	151, 169, 195, 223	320, 397, 488, 583	374, 455, 614, 786	259, 314, 461, 664	195, 249, 348, 517	100, 129, 180, 266	77, 108, 145, 208
$0 \le P < 1$	121, 139, 157, 188	357, 426, 526, 618	607, 762, 1041, 1353	677, 854, 1226, 1694	447, 596, 881, 1291	327, 442, 632, 987	167, 225, 335, 555
$1 \le P < 2$	16, 20, 26, 29	156, 173, 203, 238	530, 658, 841, 1040	901, 1213, 1734, 2248	855, 1239, 1868, 2683	753, 1022, 1496, 2268	454, 585, 880, 1320
$2 \le P < 3$			98, 112, 136, 157	205, 309, 441, 574	514, 714, 1030, 1427	587, 856, 1324, 2069	404, 616, 997, 1633
$3 \le P < 4$				**6, 6, 6, 7**	54, 74, 104, 137	112, 173, 279, 469	162, 260, 415, 686
$P \ge 4$					**2, 2, 2, 2**	**6, 6, 7, 8**	27, 43, 66, 100

(b) secondary amines

	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	77, 84, 92, 100	191, 223, 265, 317	158, 218, 304, 388	122, 159, 241, 332	79, 108, 172, 250	51, 70, 102, 141	25, 36, 54, 86
$0 \le P < 1$	67, 77, 84, 104	126, 157, 204, 266	267, 347, 486, 630	256, 364, 529, 749	297, 433, 578, 795	178, 257, 364, 538	130, 177, 271, 417
$1 \le P < 2$	16, 21, 34, 43	65, 84, 107, 142	191, 248, 357, 463	393, 535, 733, 970	579, 830, 1169, 1557	633, 876, 1208, 1697	406, 585, 815, 1171
$2 \le P < 3$		**6, 6, 6, 6**	54, 72, 108, 155	183, 252, 380, 505	460, 637, 881, 1091	584, 788, 1085, 1529	569, 750, 1019, 1459
$3 \le P < 4$				11, 14, 25, 35	99, 138, 206, 282	199, 245, 334, 520	263, 376, 513, 742
$P \ge 4$					**5, 5, 5, 6**	11, 16, 18, 26	45, 51, 69, 99

(c) carboxylic acids

	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	32, 37, 40, 42	168, 195, 227, 245	277, 335, 424, 494	212, 264, 381, 488	169, 217, 286, 387	116, 150, 215, 285	45, 66, 96, 167
$0 \le P < 1$	31, 36, 38, 43	112, 120, 146, 159	393, 481, 601, 711	514, 634, 866, 1109	391, 509, 736, 1046	333, 459, 671, 991	245, 340, 519, 822
$1 \le P < 2$		47, 54, 59, 64	151, 170, 197, 233	578, 722, 919, 1098	825, 1051, 1452, 1930	745, 1037, 1621, 2339	543, 766, 1286, 1960
$2 \le P < 3$			19, 22, 23, 27	123, 137, 159, 184	360, 463, 626, 827	571, 797, 1201, 1699	638, 872, 1414, 2098
$3 \le P < 4$				**4, 6, 6, 7**	25, 28, 31, 37	113, 131, 184, 236	184, 267, 456, 766
$P \ge 4$						**7, 7, 7, 7**	12, 16, 20, 30

(d) alcohols

	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	120, 126, 136, 154	232, 267, 329, 384	152, 193, 236, 321	103, 135, 198, 284	92, 116, 139, 190	58, 70, 93, 119	38, 52, 63, 96
$0 \le P < 1$	95, 108, 120, 129	202, 230, 280, 325	284, 360, 466, 597	229, 286, 376, 507	173, 234, 333, 454	136, 179, 239, 342	102, 130, 181, 264
$1 \le P < 2$	14, 14, 14, 15	134, 150, 173, 191	203, 242, 309, 361	337, 419, 560, 790	304, 440, 608, 867	295, 387, 521, 740	166, 222, 307, 440
$2 \le P < 3$			108, 126, 135, 150	194, 229, 302, 443	151, 222, 358, 785	223, 300, 466, 951	205, 271, 376, 638
$3 \le P < 4$				30, 30, 32, 33	46, 53, 105, 208	62, 89, 205, 525	44, 69, 170, 536
$P \ge 4$				**1, 1, 1, 1**	**1, 1, 2, 2**	15, 16, 18, 23	16, 19, 37, 60

(e) ketones

	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	32, 34, 41, 43	18, 24, 29, 33	37, 43, 56, 70	32, 38, 48, 58	25, 32, 36, 49	10, 17, 20, 30	9, 12, 20, 30
$0 \le P < 1$	23, 27, 30, 32	71, 87, 107, 117	134, 167, 214, 264	112, 147, 173, 222	84, 122, 170, 234	48, 72, 117, 178	29, 38, 58, 87
$1 \le P < 2$	24, 24, 25, 25	71, 76, 87, 96	149, 181, 230, 263	293, 372, 487, 598	309, 403, 576, 804	255, 338, 524, 742	154, 213, 328, 503
$2 \le P < 3$		13, 13, 15, 16	95, 108, 119, 146	212, 268, 352, 470	279, 359, 542, 857	383, 488, 766, 1187	331, 450 738, 1126
$3 \le P < 4$				33, 37, 39, 68	86, 116, 213, 354	219, 279, 484, 793	322, 389, 858, 1295
$P \ge 4$					10, 11, 12, 13	22, 24, 38, 116	119, 146, 272, 409

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

Table 3. continued

	(f) aldehydes						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	15, 17, 18, 20	10, 17, 20, 24	16, 19, 21, 24	38, 43, 53, 59	52, 55, 63, 76	8, 12, 19, 25	**2, 2, 4, 9**
$0 \le P < 1$	30, 34, 36, 38	87, 100, 116, 136	62, 78, 99, 124	57, 75, 100, 128	35, 48, 69, 101	18, 24, 39, 54	8, 15, 22, 33
$1 \le P < 2$	21, 23, 23, 24	61, 66, 73, 83	156, 189, 230, 268	260, 337, 423, 504	213, 298, 394, 494	130, 175, 251, 346	66, 100, 141, 215
$2 \le P < 3$		**3, 3, 3, 3**	55, 58, 68, 91	150, 187, 225, 277	230, 310, 410, 563	237, 343, 487, 693	141, 202, 323, 456
$3 \le P < 4$				19, 19, 20, 26	40, 52, 70, 89	85, 120, 184, 310	105, 189, 289, 453
$P \ge 4$					**4, 4, 4, 5**	7, 9, 11, 16	15, 19, 25, 37
	(g) boronic acids						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	**5, 5, 8, 9**	22, 26, 30, 34	96, 111, 130, 143	201, 236, 284, 337	197, 255, 329, 415	115, 156, 232, 294	62, 90, 112, 152
$0 \le P < 1$	**0, 1, 3, 3**	11, 11, 12, 13	11, 13, 18, 19	39, 47, 58, 61	93, 116, 140, 162	109, 158, 190, 256	73, 102, 150, 216
$1 \le P < 2$			**2, 3, 5, 6**	**1, 8, 9, 12**	**7, 8, 10, 10**	26, 30, 42, 45	28, 57, 81, 103
$2 \le P < 3$					**1, 1, 2, 3**	**1, 1, 2, 2**	5, 7, 10, 13
$3 \le P < 4$						**1, 1, 1, 1**	
$P \ge 4$							**1, 1, 1, 1**
	(h) acyl halides						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$		**1, 1, 1, 1**			**2, 2, 2, 2**		
$0 \le P < 1$	**8, 8, 8, 9**	10, 10, 13, 13	15, 17, 21, 26	11, 14, 17, 17	**4, 5, 6, 16**	**2, 4,** 7, 7	**1,1, 2, 2**
$1 \le P < 2$	**1, 1, 1, 1**	23, 24, 28, 30	35, 43, 55, 56	44, 53, 65, 76	50, 58, 73, 95	21, 24, 29, 46	8, 10, 15, 23
$2 \le P < 3$		**1, 1, 1, 1**	18, 19, 21, 22	59, 67, 77, 83	100, 119, 151, 193	95, 123, 160, 219	41, 55, 73 101
$3 \le P < 4$				**2, 2, 2, 2**	16, 19, 24, 34	33, 40, 68, 99	53, 80, 115, 215
$P \ge 4$						**5, 6, 6, 6**	11, 12, 14, 25
	(i) sulfonyl chlorides						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$		**1, 1, 1, 1**	**2, 2, 2, 2**	**1, 1, 1, 2**	**1, 4, 6, 6**	**2, 3, 4, 5**	**0, 0, 2, 3**
$0 \le P < 1$		**3, 3, 3, 3**	**6, 6, 6, 9**	10, 11, 17, 22	12, 14, 26, 28	7, 8, 14, 17	7, 12, 18, 32
$1 \le P < 2$				7, 9, 12, 15	29, 34, 38, 52	50, 62, 82, 106	45, 60, 97, 145
$2 \le P < 3$					**2, 2, 2, 2**	29, 33, 40, 48	66, 81, 111, 154
$3 \le P < 4$							**6, 8, 8, 10**
$P \ge 4$							
	(j) isocyanates						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$	**0, 0, 0, 1**	**0, 0, 1, 3**	**2, 2, 3, 5**	**0, 0, 0, 2**	**1, 1, 2, 3**		
$0 \le P < 1$	**4, 6, 7, 9**	**3, 3, 6, 9**	**4, 6, 7, 14**	**5, 7, 8, 9**	**1, 2, 4, 4**	**1, 3, 3, 6**	**0, 0, 2, 2**
$1 \le P < 2$	**3, 3, 3, 3**	5, 7, 12, 16	21, 25, 29, 33	19, 28, 33, 40	25, 37, 46, 52	7, 12, 19, 20	**1, 4, 6, 6**
$2 \le P < 3$			12, 14, 14, 16	36, 44, 60, 65	31, 42, 49, 56	8, 18, 28, 37	2, 7, 11, 17
$3 \le P < 4$				**0, 0, 1, 3**	4, 11, 14, 16	8, 16, 22, 23	0, 6, 10, 11
$P \ge 4$						**1, 1, 1, 2**	**3, 3, 5, 5**

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

Table 3. continued

	(k) isothiocyanates						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$					**1, 1, 1, 1**		
$0 \le P < 1$	**1, 2, 2, 2**	**2, 3, 3, 3**	**2, 5, 6, 6**	**2, 2, 2, 3**	**1, 1, 2, 2**	**0, 2, 3, 4**	**0, 0, 0, 1**
$1 \le P < 2$	**3, 3, 3, 3**	**8, 10, 10, 10**	10, 13, 15, 19	15, 22, 26, 33	10, 11, 17, 22	**7, 11, 14, 15**	**0, 2, 3, 6**
$2 \le P < 3$			15, 15, 18, 20	37, 42, 53, 60	49, 56, 72, 84	14, 18, 23, 35	3, 6, 7, 19
$3 \le P < 4$				16, 16, 20, 25	33, 38, 46, 54	15, 22, 29, 38	21, 27, 37, 58
$P \ge 4$					**1, 1, 1, 1**	**3, 4, 7, 8**	13, 17, 19, 20
	(l) azides						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$					**0, 0, 0, 2**	**0, 1, 1, 1**	**0, 0, 1, 2**
$0 \le P < 1$	**0, 1, 1, 1**	**1, 2, 2, 6**	**3, 3, 3, 5**	**1, 3, 4, 5**	**1, 1, 2, 4**	**1, 3, 4, 4**	**1, 1, 1, 2**
$1 \le P < 2$	**0, 0, 0, 1**	**1, 1, 1, 2**	7, 7, 10, 13	**5, 5, 6, 6**	**3, 3, 5, 9**	**0, 0, 1, 3**	**1, 2, 2, 4**
$2 \le P < 3$		**0, 0, 0, 2**	2, 2, 6, 18	5, 5, 7, 16	3, 4, 9, 15	**1, 2, 4, 7**	5, 6, 8, 11
$3 \le P < 4$			**0, 0, 0, 9**	1, 1, 2, 14	**1, 4, 6, 8**	**1, 1, 5, 5**	**1, 2, 4, 5**
$P \ge 4$						**0, 0,0, 1**	**0, 0, 0, 1**
	(m) chloroformates						
	$0 \le W < 100$	$100 \le W < 125$	$125 \le W < 150$	$150 \le W < 175$	$175 \le W < 200$	$200 \le W < 225$	$225 \le W < 250$
$P < 0$							
$0 \le P < 1$	**1, 1, 1, 1**						
$1 \le P < 2$		**6, 6, 6, 6**	**3, 4, 7, 7**	**1, 3, 3, 4**	**1, 1, 1, 1**		
$2 \le P < 3$			**4, 4, 4, 4**	**9, 9, 9, 9**	**4, 4, 4, 5**	**6, 7, 7, 7**	
$3 \le P < 4$					**5, 5, 5, 5**	**3, 5, 6, 6**	**3, 4, 4, 4**
$P \ge 4$						**1, 2, 2, 2**	**1, 1, 1, 1**

[a]Different cutoffs are separated by comma (25, 50, 100, and 150 USD per gram). The content of cells with ≤10 building blocks are bold.

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

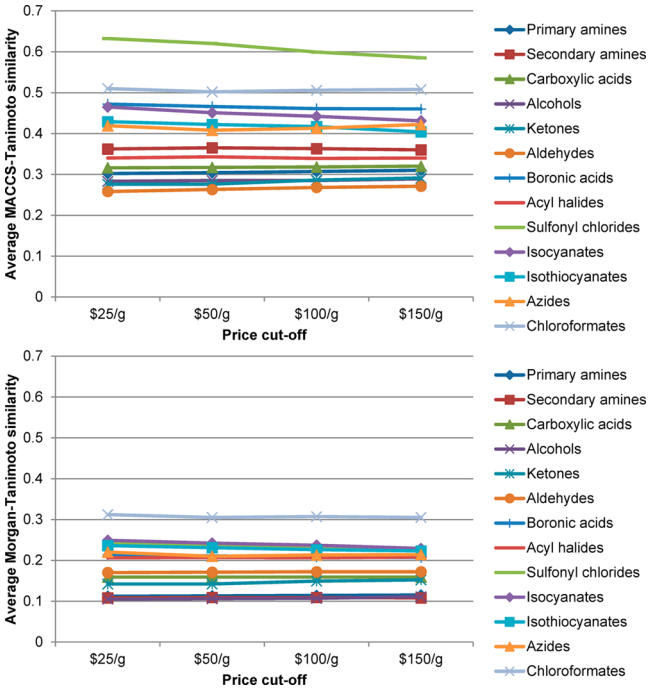


Figure 1. Average intraclass MACCS and Morgan-Tanimoto similarities using different price cutoffs (25, 50, 100, and 150 USD per gram).

compounds (only 161 building blocks available with $150/gram cutoff). However, slightly increasing the molecular weight cutoff above 150 Da to 150−175 Da range would allow consideration of much larger chemical space (701 building blocks available with $150/gram cutoff).

More expensive building blocks can be justified in the design process if they increase the diversity of the library and thus enable the exploration of novel chemical space. The new building blocks that were available after increasing the price cutoff were compared with 2D-fingerprints to the lower threshold ones as described in the Experimental Procedures. The average similarity measured both with MACCS- and Morgan-Tanimoto stays similar using different price cutoffs with almost all reagent classes (Figure 1). In most cases, expanding the selection by considering more expensive building blocks brings in compounds of similar, or even slightly lower, average similarity value. In other words, at every price range there is an equally diverse selection of building blocks available. This effect is most clearly observed with the sulfonyl chlorides where the MACCS-Tanimoto similarity goes down with increasing price cutoffs from 0.632 to 0.585. Considering the more expensive building blocks does increase the number of different functional groups attached to the sulfonyl chlorides, but this analysis shows that one should first exhaust the cheaper chemical space before buying more expensive compounds.

Cost does make a difference in the percentage of building blocks allowing for three-dimensional structural projection, as opposed to a "flat" nature, for a few classes. There is not a generally accepted descriptor for the 3D-likeness of molecules. In this study, commonly used metric of Fsp3 was selected.[8] Fsp3 is defined as the number of sp^3 hybridized carbons divided by the total carbon count. Analysis of the 3D-likeness versus price of aldehydes, acyl halides, sulfonyl chlorides and isocyanates shows that more expensive building blocks in these subtypes have larger Fsp3 (Figure 2). For example, building blocks costing less than 25 USD per gram exhibit median Fsp3 values of less than 0.20. However, for other reagents classes, increasing the price will not enable access to building blocks with significantly increased Fsp3. Azides and chloroformates produce very striking curves, but the large median Fsp3 differences between the price ranges are caused by the low number of compounds (for example, only 16 building blocks exist in the price group 25−50 USD) and it is difficult to make any conclusions for these reagent classes.

The numbers provided in this paper most likely overestimate the number of actual building blocks available for library synthesis and they should be taken as only rough guidelines. Not every building block is available from vendors' listings, and we do not take into account other undesirable features that would usually exclude certain structures from a new library, such as compounds prone to pan-assay interference.[6,7]

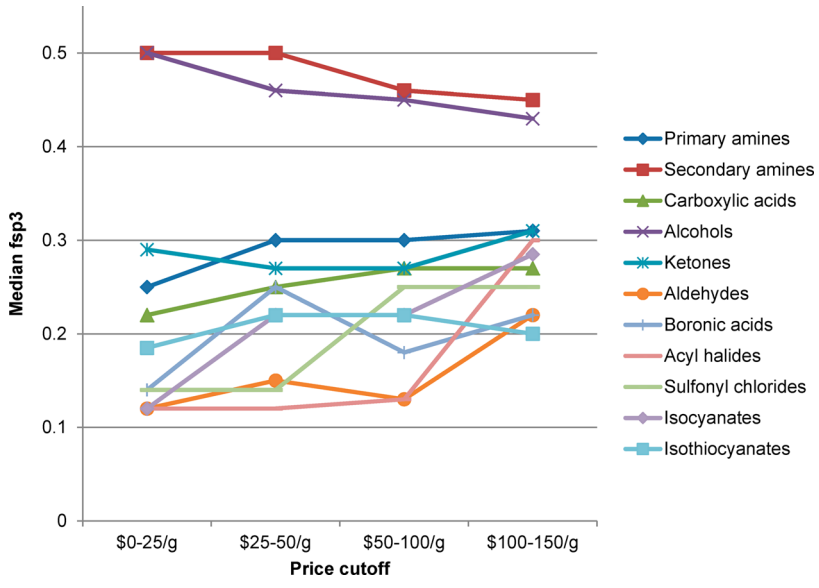


Figure 2. Median Fsp3 of the building blocks when using different price ranges (azides and chloroformates are excluded for clarity).

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

CONCLUSIONS

The chemical space of combinatorial libraries is practically infinite,[16] but the chemical space of building blocks and the number of the chemical reactions for decorating these scaffolds is indeed limited, and budgetary constrains invariably impinge on practical choices. A significant number of commercial building blocks are too expensive for large-scale combinatorial library synthesis in most settings, and certain reagent classes such chloroformates are especially limited. Other classes, such as primary or secondary amines, are widely available at reasonable prices. For most classes, however, cost constraints do not severely limit diversity as assessed by 2D fingerprint-type analysis, but do impact the property of three-dimensional structural projection. Although the relationship between molecular weight/logP and price of a building block is not straightforward, tables linking these properties are presented that can assist in combinatorial library design.

EXPERIMENTAL PROCEDURES

The data set was extracted from eMolecules Plus building block collection database using combination of JChem[17] and RDKit[18] Nodes implemented in KNIME.[19] The database contained 963 960 compounds from 121 vendors. First, molecules were imported in SMILES format and salts stripped. Then, molecules that had other atoms than carbon, nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, bromine, iodine, hydrogen, or boron were excluded from the data set. Building blocks that had molecular weight higher than 250 Da were excluded to avoid high molecular weight of the final library compounds. The cheapest price per gram in USD was picked for each of the remaining molecules from the all possible vendors and package sizes stored in the database. As the aim of this study was to study the cheaper end of the building blocks' chemical space, a price cutoff 1000 USD per gram was applied after which 413 994 reagents were left.

The reagent classes were extracted using either JChem's Chemical Terms-node (matchCount-function) or MolSearch-node if there was no predefined chemical term implemented in the software for a reagent class. Only one query functional group per building block was allowed, but other functional groups per building blocks were not controlled. This means that the same amino acids, for example, were included both in primary amine and carboxylic acid groups, but on the other hand no alcohol building block could have two hydroxyl-groups attached.

Chemical similarity was accessed by two commonly used 2d-metrics: MACCS- and Morgan-fingerprints (similar to the ECFP4-fingerprints). These two fingerprints are very different and measure the similarity between two molecules from a different viewpoint. The fingerprints and similarity calculations were done using ChemFP.[20] Average internal chemical similarity was computed by first comparing all compounds inside a library to each other and taking the average similarity for each of the compounds. The average of these averages was then used as the metric for internal diversity (the smaller the number is, the more diverse library is).

Fsp3-values were calculated using JChem cxcalc. All statistics were calculated with combination of Python/NumPy-scripts and R.[21]

AUTHOR INFORMATION

Corresponding Author

*E-mail: kalliokoski@lead-discovery.de. Tel: +49 (0) 231-97427056.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Uwe Koch and Peter Nussbaumer are thanked for their comments on the manuscript. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115489, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

ABBREVIATIONS

BB, building block; ELF, European lead factory; Fsp3, fraction sp3 (the number of sp3 hybridized carbons divided by the total carbon atom count); GLARE, global library assessment of reagents

REFERENCES

(1) Tiebes, D. Combinatorial Chemistry. In *Combinatorial Chemistry: Synthesis, Analysis, Screening*; Jung, G., Ed.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 1999; pp 1−34.

(2) Bannwarth, W.; Hinzen, B. *Combinatorial Chemistry: From Theory to Application*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2005; pp XI−XX.

(3) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.

(4) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(5) Truchon, J. F.; Bayly, C. I. GLARE: a new approach for filtering large reagent lists in combinatorial library design using product properties. *J. Chem. Inf. Model.* **2006**, *46*, 1536−1548.

(6) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(7) Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **2012**, *55*, 9763−9772.

(8) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752−6756.

(9) Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *J. Chem. Inf. Model.* **2012**, *52*, 2516−2525.

(10) Walters, W. P.; Green, J.; Weiss, J. R.; Murcko, M. A. What do medicinal chemists actually make? A 50-year retrospective. *J. Med. Chem.* **2011**, *54*, 6405−6416.

(11) European Lead Factory. https://www.europeanleadfactory.eu/ (accessed Jan 2015).

(12) Accelrys Available Chemical Directory (ACD). http://accelrys.com/products/pdf/available-chemicals-directory.pdf.

(13) eMolecules, Inc. Emolecules Plus database. http://www.emolecules.com (accessed Jan 2015).

(14) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(15) Brown, D. B.; Gagnon, M. M.; Boström, J. Understanding Our Love Affair with p-Chlorophenyl: Present Day Implications from Historical Biases of Reagent Selection. *J. Med. Chem.* **2015**, *58*, 2390.

(16) Lowe, D. Chemical space is big. Really big. *MedChemComm* **2015**, *6*, 12.

(17) *JChem*; ChemAxon Kft.: Budapest, Hungary; http://www.chemaxon.com.

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607

(18) RDKit: Open-source cheminformatics. http://www.rdkit.org.
(19) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz informationminer: version 2.0 andbeyond. *SIGKDD Explor. Newsl.* **2009**, *11*, 26−31.
(20) Dalke, A. The FPS fingerprint format and chemfp toolkit. *J. Cheminf.* **2013**, *5*, P36.
(21) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria; http://www.R-project.org.

DOI: 10.1021/acscombsci.5b00063
ACS Comb. Sci. 2015, 17, 600−607