Electronic Supplementary Material (ESI) for Medicinal Chemistry Communications


**Supporting Information**

**Title:** Structural Enrichment of HTS Compounds from Available Commercial Libraries.

**Authors:** Tetyana Petrova, Alexander Chuprina, Raman Parkesh, and Alexei Pushechnikov*

**List of contents**

## Methods

***Data preparation**:* Instant JChem was used for structure database management, search and prediction,[1] and MySQL[2] database server was used to store and manage supplier information, structures, properties and structural features. All libraries were standardized prior to the full analysis. Inorganic compounds and structures containing metals were excluded from further consideration. Radicals, solvents and salt data (counterions) were removed from the structure fields and charges were neutralized. Nitro-, sulfoxide-, and nitroxide-groups were transformed into the charge-separated ones, the information on the absolute configuration of stereocenters was removed, and the structures were converted into the canonic tautomers. ChemAxon Standardizer was used for structure canonicalization and transformation.[3] The duplicate structures were removed from the individual supplier libraries to give the collection of unique compounds using CheD program.[4]

***Structural filtering and calculation of properties**:* The structural filters were applied as substructure query strings for the chemical terms expression represented as SMARTS[5] using JChem Evaluator.[3] ChemAxon Calculator Plugins were used for physicochemical properties (ClogP, MW, PSA, etc.) prediction and calculation.[6] Resulting parameters were further used for estimation of library composition, filtering 'drug-like' compounds and calculation of descriptors of molecular complexity.

Measure of the carbon bond saturation (Fsp3, Lovering et al.[7]) is defined as the number of sp3 hybridized carbons divided by the total carbon count.

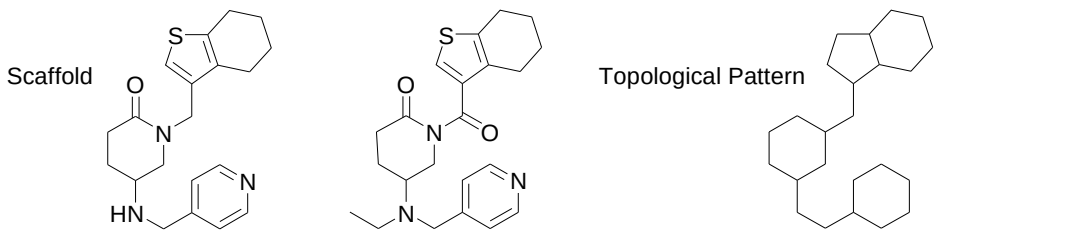
$$\mathrm{Fsp}^3 = \frac{Ncarbon_{sp^3}}{Ncarbon_{total}}$$

Descriptor of the structural complexity (fMF, Yang et al.[8]) is defined as the number of heavy atoms in the molecular framework (MF) divided by the total number of heavy atoms in the molecule:

$$f_{\mathrm{MF}} = \frac{Nheavy_{MF}}{Nheavy_{total}}$$

The distribution of benzenoid rings was calculated, since the nature of aromatic rings was found to be an important factor in the compound developability.[9-11]

***Diversity assessment**:* The derived scaffolds (ChemAxon Calculator[6] topology analysis plugin function) in each database were represented by extended connectivity fingerprint with bond diameter 4 (ECFP_4) and 1024-bit set size.[12] This descriptor was chosen as the most suitable fingerprint type for the purpose of our analysis because of its optimal combination of computational time and accuracy,[13, 14] as well as its highest apparent potential for the scaffold hopping.[15] The structural diversity of individual compounds was estimated via clustering around topological patterns as centroids using Bemis-Murcko framework based clustering.[16] ChemAxon JKlustor tools were used for ECFP calculation, clustering and diversity analysis of chemical sets.[3]



***Similarity calculation**:* The derived scaffolds of libraries after the structural enrichment were compared via the pair-wise fingerprint distance calculation. The average distances were based on ECFP_4 structural level and derived via "compr" function of ChemAxon JKlustor module. The ECFP fingerprints were calculated evoking ChemAxon command "generatemd" with parameters: "-f 1024 -n 4 -b 2 -k ECFP -2 –g". The fingerprint distances between compounds of two libraries were calculated evoking ChemAxon command "compr". The algorithm applies nearest neighbour searching that finds molecules similar to the query object. Then the directional average shortest distance between two sets of scaffolds was calculated through the normalization of the sum of "minD" values to the size of a respective library. The asymmetrical similarity for a pair of libraries was expressed as a pair of the corresponding similarity indexes by subtracting each average distance from 1, e.g. if A(N) and B(M) have N>M then:

*S*(A,B) - estimation of a redundancy degree of the set A relatively to the set B
"compr -f 1024 -t 0.25 -z -i B A"
$S(A,B) = 1 - \Sigma_i(minD_i)/N$

*S*(B,A) - estimation of a congruency degree of the set B relatively to a continuity of subsets of the set A
"compr -f 1024 -t 0.25 -z -i A B"
$S(B,A) = 1 - \Sigma_i(minD_i)/M$

## SMARTS filters

List of SMARTS strings for removal compounds with undesirable functional groups/substructures from screening libraries.

```
################################################################################
#count rules
################################################################################
6	Fluorines	F
3	Halogens_Cl_Br_I	[Cl,Br,I]
1	Naphthalenes	c12ccccc1cccc2
1	Nitro_Group	N(~[OD1])~[OD1]
################################################################################
#no occurrence rules
################################################################################
0	Acid_anhydrides	[CX3;#6](=[OX1])[#8X2][CX3;#6](=[OX1])
0	Acid_halides	[S,C](=[OX1,SX1])[F,Br,Cl,I]
0	Acrylonitriles_1	N#CC=C
0	Acrylonitriles_2	[!#7;!#8;!#16][CX3;CH]=[CX3;CH](C#N)
0	Acylazides	[NX1]~[NX2]~[NX2]C=[OX1]
0	Acylcyanides	N#CC(=[OX1])
0	Acylhydrazides	[N;R0][N;R0]C(=[OX1])
0	Aldehydes	[CX3H1](=[OX1])[#6]
0	Aliph_esters	[C;R0][CX3](=[OX1])[OX2][C;R0]
0	Aliph_ketones	[C;R0][CX3](=[OX1])[C;R0]
0	Aliphatic_methylene_chains_7_or_longer	[CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0]
0	Alkyl_halides	[Br,Cl,I][$([CX4;R0]),$([CX3;R0])][#1]
0	Aminooxy_oxo	[#7][#8][#6,#16]=[OX1]
0	Anthracenes	c12ccccc1cc3ccccc3c2
0	Aromatic_azides	[NX1]~[NX2]~[NX2]c
0	Azoalkanals	[N;R0]=[N;R0]CC=[OX1]
0	Azocyanamides	[N;R0]=[N;R0]C#N
0	Benzylic_quaternary_N	c[C;R0][NX4+]
0	Beta_carbonyl_quaternary_N	C(=[OX1])[#6][NX4+]
0	Carbodiimides	[NX2]=C=[NX2]
0	Cations_C_Cl_I_P_S	[#6+,#17+,#53+,#15+,#16+]
```

```
0	Compounds_with_4_more_acidic_groups	[C,S,P](=O)[OH].[C,S,P](=O)[OH].[C,S,P](=O)[OH].[C,S,P](=O)[OH]
0	Crown_ethers	[O;R1][C;R1][C;R1][O;R1][C;R1][C;R1][O;R1]
0	Cyanohydrines	[NX1]#CC[OX2H]
0	Cyanophosphonates	P([OX2]CC)([OX2]CC)(=[OX1])C#N
0	Di_or_Triphosphates	P(=[OX1])([OX2H])[OX2]P(=[OX1])[OX2H]
0	Dicarbonyl_groups	[CX3;R0](=[OX1])[CX3;R0](=[OX1])
0	Disulfides	[#16][#16]
0	Enamines	[#6][CX3](!@N)=[CX3][#6]
0	Epoxides_Thioepoxides_Aziridines	C1[O,S,N]C1
0	Halopyrimidines	[F,Br,Cl,I]c1[#7X2]ccc[#7X2]1
0	Hexanes	[CD1][CD2][CD2][CD2][CD2][CX4;CH2,CH3]
0	HOBT_esters	C(=[OX1])Onnn
0	Hydrazines	[N;R0]-[N;R0]
0	Hydrazothiourea	N=NC(=S)N
0	Hydroxylamines1	[N;R0;!$([N+])]-[OX2;R0]
0	Hydroxylamines2	[NX4+;R0]([#1])-[OX1-;R0]
0	Imines	[#1,#6][CX3](=[N;R0])([#1,#6])
0	Isocyanates_Isothiocyanates	[$([NX1-]),$([NX2])]=[CX2]=[S,O]
0	Isonitrile	[NX2+]#[CX1-]
0	Lawesson_reagent_derivatives	P(=S)(S)S
0	Macrocycles	[r10,r11,r12,r13,r14,r15,r16,r17,r18,r19,r20,r21,r22,23,r24,r25,r26,r27,r28,r29,r30,r31,r32,r33,r34,r35,r36,r37,r38,r39,r40]
0	Malonolyden_Dinitriles	C=C(C#N)(C#N)
0	Malonolyden_Nitrile_Esters	[CX3;CH,CH2]=[CX3](C#N)[C;R0](=[OX1])[!#7;!#8]
0	Michael_Acceptors	[CH1,CH2]=CC=[OX1]
0	N_halides	[#7][Cl,Br,I]
0	N_oxides	[$([#7+][OX1-]),$([#7v5]=[OX1]);!$([#7](~[OD1])~[OD1])]
0	N_S_heteroat_bond	[N;R0]-[S;R0;!v6]
0	Nitroso_groups	[NX2]=[OX1]
0	No_occur_atoms	[!#1;!#6;!#7;!#8;!#9;!#11;!#12;!#15;!#16;!#17;!#19;!#20;!#35;!#53]
0	P_or_S_Halides	[P,S][Cl,Br,F,I]
0	Paranitrophenyl_esters	[CX3;R0](=[OX1])[OX2]c1ccc([$([NX3](=O)=O),$([NX3+](=O)[O-])])cc1
0	Pentafluorophenyl	c1c(F)c(F)c(F)c(F)c1(F)
0	Perchlorates	OCl(O)(O)(O)
0	Perhaloketones	[C;R0][CX3](=[OX1])[CX4]([F,Cl,Br,I])([F,Cl,Br,I])[F,Cl,Br,I]
0	Peroxides	[OX2][OX2]
0	Phenantrenes	c12ccccc1ccc3ccccc23
0	Phosphenes	cPc
0	Phosphonate_ester	[#6][OX2]P(=[OX1])[OX2][#6]
0	Phosphoramides	NP(=[N,O,S])(N)N
0	Phosphoranes	C=P
0	Phosphorus_3	[Pv3]
0	Polyenes_acycl	[CX3;R0]=[CX3;R0][CX3;R0]=[CX3;R0][CX3;R0]=[CX3;R0][CX3;R0]=[CX3;R0]
0	Quinones	[OX1]=[#6]1[#6]~[#6][#6](=[OX1])[#6]~[#6]1
0	Sulfate_esters	[#6][OX2]S(=[OX1])(=[OX1])[OX2][#6]
0	Sulfonates	[#6]S(=[OX1])(=[OX1])[OX2][#6]
0	Sulfonyl_cyanides	S(=[OX1])(=[OX1])C#N
0	Sulfonyl_halides	S(=[OX1])(=[OX1])[F,Cl,Br,I]
0	Thiocyanates	SC#[NX1]
0	Thioesters	[#6][O,S;R0][C;R0](=S)
0	Thioesters_2	[#6][S;R0][C;R0](=[OX1])
0	Triacyloximes	C(=[OX1])[NX3](C(=[OX1]))[OX2]C(=[OX1])
0	Triflates	[OX2]S(=[OX1])(=[OX1])C(F)(F)F
#############################################################################
#  PAINS
#############################################################################
0	Anil_di_alk_A	[CH2,CH3;CX4][NX3;R0]([CH2,CH3;CX4])c1cc([$([#1]),$([CH2,CH3;CX4]),$(O[CX4;CH2][CH2,CH3;CX4])])c([#7])c(H)c1
0	Anil_di_alk_B	[CX4][NX3]([CX4])c1ccc([CX3]=[CX3])cc1
0	Anil_di_alk_C	[CX4]N([$([#1]),$([CX4])])c1ccc([OX2][CX4])cc1
0	Anil_di_Alk_D	[CX4][NX3]([CX4])c1ccc([CX4][$([OX2H]),$([CX3]=[CX3][#1]),$([NX3][CX4])])cc1
```

0 Anil_di_Alk_E [CX4;CH2,CH3][NX3](C[#1])c1c([#1])c([$([#1]),$([CX4;CH2,CH3])])c(c([#1])c1([#1]))C([#1])[$([#1]),$(C([#1]))]
0 Azo_A [N;R0]=[#7]
0 Ene_five_het_A C1(=[CX3])C(=[OX1])[#7,#8,#15,#16]N=C1
0 Ene_rhod_A C1(=[SX1])NC(=[OX1])C(=[#6])S1
0 Ene_six_het_A [#7,#8,#15,#16]~C1~[#7,#8,#15,#16]~C[#7,#8,#15,#16][CX3](=[#7,#8,#15,#16])[CX3]1(=[CH,CH2;R0])
0 Hzone_phenol_A [OX2H]c1c([CX3]=[NX2;R0][NX3;R0])cccc1
0 Hzone_phenol_B [OX2H]c1ccc([CX3]=[NX2;R0][NX3;R0])cc1
0 Imine_one_A [#6][#6](=[#7,#8,#15,#16;R0])[#6](=[#7,#8,#15,#16;R0])[$([#6]),$(S(=[OX1])=[OX1])]
0 Imine_one_isatin [OX1]=C1[#7]c2ccccc2C1=[NX2;R0][#7]
0 Indol_3yl_alk N1([$([#1]),$([CX4]([#1])[#1])])C([$([CX4]([#1])[#1]),$(C([#1])N),$([CD3](C([#1])[#1])N([#1])C([#1])[#1]),$([CD3](C([#1])[#1])[CD2]N([#1])C([#1])[#1]),$([#6]=[#7,#8,#15,#16]),$([#6]:[#7,#8,#15,#16])])=C([CX4]H)c2c(H)cccc12
0 Mannich_A [OX2H]c1c([CX4][#7])cccc1
0 Pyrrole_A N1(c2[!#1]cccc2)C([CX4])=CC(H)=C1[CX4]
0 Quinone_A [#7,#8,#15,#16]=[#6]1[$([#6]=[#6]),$([#6]:[#6])][#6](=[#7,#8,#15,#16])[$([#6]=[#6]),$([#6]:[#6])]1
0 Thiaz_ene_A [#7X2;R0]=[CX3]1[SX2][CX3]([$([#1]),$([CX4]([#1])[#1]),$([CX3]=[OX1])])=[CX3]([$([CX4]([#1])[#1]),$([#6]:[#6])])[NX3]1([$([#1]),$([#6][#1]),$([#6]:[#6])])
##############################################################################

**Table S1** Information about the commercially available libraries and the NCI open database (gathered in February 2011)

| Supplier [a] | Supplier website | Number of compounds | Number of exclusive compounds |
|---|---|---|---|
| Abamachem | http://www.abamachem.com | 1,433,640 | 1,394,618 |
| AMRI | http://www.amriglobal.com | 253,405 | 250,560 |
| Alinda Chemical Ltd. [b] | http://www.alinda.ru | 256,282 | 14,339 |
| AllLab | http://www.albchemical.com | 19,517 | 4,429 |
| AnalytiCon Discovery [b] | http://www.ac-discovery.com | 24,237 | 23,986 |
| Aronis | http://www.aronis.ru | 22,927 | 134 |
| ASDI | http://www.asdi.net | 108,191 | 10,231 |
| ASINEX | http://www.asinex.com | 412,485 | 230,623 |
| BCHResearch | http://www.bchresearch.com | 1,499,151 | 1,449,359 |
| ChemBridge | http://www.chembridge.com | 634,160 | 264,181 |
| ChemDiv | http://www.chemdiv.com | 825,535 | 482,855 |
| Chemical Block | http://www.chemblock.com | 121,063 | 1,364 |
| EMC microcollections [b] | http://www.microcollections.de | 26,570 | 26,335 |
| Enamine | http://www.enamine.net | 1,636,699 | 1,048,876 |
| FCHGroup | http://www.fchgroup.net | 1,455,286 | 1,443,239 |
| Focus Synthesis | http://www.focussynthesis.com | 1,003 | 651 |
| InterBioScreen | http://www.ibscreen.com | 450,740 | 92,663 |
| Intermed Ltd | http://www.intermed.dn.ua | 32,106 | 3,204 |
| Key Organics | http://www.keyorganics.ltd.uk | 52,043 | 40,077 |
| Life Chemicals | http://www.lifechemicals.com | 329,403 | 185,542 |
| Maybridge | http://www.maybridge.com | 60,926 | 39,336 |
| Menai Organics | http://www.menaiorganics.co.uk | 3,596 | 1,370 |
| Molecular Diversity Preservation Intl [b] | http://www.mdpi.org | 18,358 | 13,046 |
| Nanosyn | http://www.nanosyn.com | 62,579 | 17,556 |
| Otava | http://www.otavachemicals.com | 90,191 | 19,824 |
| Peakdale Molecular [b] | http://www.peakdale.co.uk | 15,220 | 14,171 |
| Princeton BioMolecular Research [b] | http://www.princetonbio.com | 608,389 | 63,098 |
| Selena Chem | http://www.selenachem.com | 1,537,890 | 1,492,769 |
| Sigma Aldrich | http://www.sigmaaldrich.com | 165,860 | 72,782 |
| Specs | http://www.specs.net | 198,771 | 47,560 |
| Synthon-Lab | http://www.synthon-lab.com | 46,272 | 6,307 |
| TimTec | http://www.timtec.net | 196,086 | 15,383 |
| TosLab | http://www.toslab.com | 16,215 | 8,506 |
| UORSY | http://www.ukrorgsynth.com | 1,525,625 | 1,022,292 |
| Vitas M Lab | http://www.vitasmlab.com | 924,849 | 51,641 |
| Zelinsky Institute | http://www.zelinsky.com | 299,835 | 39,914 |
| NCI (*ver.* 2.1) | http://cactus.nci.nih.gov | 231,346 | 190,341 |
| **Total** | | 15,596,451 | 10,083,162 |
| **Total unique** | | 11,870,691 | 10,083,162 |

[a] In other Figures and Tables the commercial suppliers are ranked in ascending order of database size for the convenience in data representation and analysis. For further comparisons the NCI database is outlined as a reference.
[b] Abbreviations used: Alinda Chemical Ltd. - Alinda; Analyticon Discovery – AnalytiCon; EMC microcollections – EMC; Molecular Diversity Preservation Intl – MDPI; Peakdale Molecular – Peakdale; Princeton BioMolecular Research - PrincetonBio, respectively.

**Table S2** Contributions of individual substructure filters for filtering out undesirable compounds from screening libraries (11,870,691 compounds).

| Name of SMARTS filter | Number of failed compounds | Name of SMARTS filter | Number of failed compounds |
|---|---|---|---|
| # count rules | | Malonolyden_Nitrile_Esters | 3,468 |
| Fluorines | 5,919 | Michael_Acceptors | 381,488 |
| Halogens_Cl_Br_I | 5,173 | N_halides | 191 |
| Naphthalenes | 4,865 | N_oxides | 9,007 |
| Nitro_Group | 30,520 | N_S_heteroat_bond | 671 |
| # no occurrence rules | | Nitroso_groups | 1,755 |
| Acid_anhydrides | 1,255 | No_occur_atoms | 13,112 |
| Acid_halides | 4,611 | P_or_S_Halides | 3,146 |
| Acrylonitriles_1 | 70,837 | Paranitrophenyl_esters | 1,473 |
| Acrylonitriles_2 | 92 | Pentafluorophenyl | 3,516 |
| Acylazides | 40 | Perchlorates | 46 |
| Acylcyanides | 51 | Perhaloketones | 975 |
| Acylhydrazides | 386,663 | Peroxides | 475 |
| Aldehydes | 18,529 | Phenantrenes | 1,828 |
| Aliph_esters | 739,782 | Phosphenes | 3,118 |
| Aliph_ketones | 26,823 | Phosphonate_ester | 4,852 |
| Aliphatic_methylene_chains_7_or_longer | 27,659 | Phosphoramides | 549 |
| Alkyl_halides | 27,615 | Phosphoranes | 310 |
| Aminooxy_oxo | 15,548 | Phosphorus_3 | 849 |
| Anthracenes | 2,537 | Polyenes_acycl | 83 |
| Aromatic_azides | 501 | Quinones | 13,860 |
| Azoalkanals | 965 | Sulfate_esters | 35 |
| Azocyanamides | 25 | Sulfonates | 31,276 |
| Benzylic_quaternary_N | 868 | Sulfonyl_cyanides | 2 |
| Beta_carbonyl_quaternary_N | 730 | Sulfonyl_halides | 2,635 |
| Carbodiimides | 41 | Thiocyanates | 810 |
| Cations_C_Cl_I_P_S | 2,034 | Thioesters | 4,364 |
| Compounds_with_4_more_acidic_groups | 630 | Thioesters_2 | 2,610 |
| Crown_ethers | 619 | Triacyloximes | 432 |
| Cyanohydrines | 119 | Triflates | 111 |
| Cyanophosphonates | 1 | # PAINS | |
| Di_or_Triphosphates | 210 | Anil_di_alk_A | 23,161 |
| Dicarbonyl_groups | 48,782 | Anil_di_alk_B | 5,323 |
| Disulfides | 4,644 | Anil_di_alk_C | 26,058 |
| Enamines | 21,860 | Anil_di_Alk_D | 7,864 |
| Epoxides_Thioepoxides_Aziridines | 5,479 | Anil_di_Alk_E | 46,869 |
| Halopyrimidines | 1,400 | Azo_A | 12,063 |
| Hexanes | 44,069 | Ene_five_het_A | 5,049 |
| HOBT_esters | 5 | Ene_rhod_A | 29,851 |
| Hydrazines | 457,969 | Ene_six_het_A | 29,196 |
| Hydrazothiourea | 52 | Hzone_phenol_A | 14,123 |
| Hydroxylamines1 | 44,960 | Hzone_phenol_B | 9,276 |
| Hydroxylamines2 | 9 | Imine_one_A | 5,655 |
| Imines | 52,999 | Imine_one_isatin | 8,619 |
| Isocyanates_Isothiocyanates | 1,384 | Indol_3yl_alk | 37,584 |
| Isonitrile | 167 | Mannich_A | 17,421 |
| Lawesson_reagent_derivatives | 27 | Pyrrole_A | 23,768 |
| Macrocycles | 3,758 | Quinone_A | 7 |
| Malonolyden_Dinitriles | 2,771 | Thiaz_ene_A | 2,986 |

**Table S3** Analysis of the distribution of 'drug-like' properties of compounds.

| Supplier | Number of compounds | *SMARTS filter* | | *Lipinski rule* | |
|---|---|---|---|---|---|
| | | **No. of compounds passed** | **%** | **No. of compounds passed** | **%** |
| Focus Synthesis | 1,003 | 576 | 57.4 | 859 | 85.6 |
| Menai Organics | 3,596 | 2,008 | 55.8 | 3,027 | 84.2 |
| Peakdale | 15,220 | 14,224 | 93.5 | 12,786 | 84.0 |
| TosLab | 16,215 | 11,068 | 68.3 | 10,435 | 64.4 |
| MDPI | 18,358 | 10,004 | 54.5 | 13,239 | 72.1 |
| AllLab | 19,517 | 14,554 | 74.6 | 14,806 | 75.9 |
| Aronis | 22,927 | 13,614 | 59.4 | 18,875 | 82.3 |
| AnalytiCon | 24,237 | 21,182 | 87.4 | 18,191 | 75.1 |
| EMC | 26,570 | 24,310 | 91.5 | 18,479 | 69.5 |
| Intermed Ltd | 32,106 | 20,472 | 63.8 | 20,459 | 63.7 |
| Synthon-Lab | 46,272 | 19,685 | 42.5 | 35,164 | 76.0 |
| Key Organics | 52,043 | 36,472 | 70.1 | 41,777 | 80.3 |
| Maybridge | 60,926 | 41,677 | 68.4 | 49,456 | 81.2 |
| Nanosyn | 62,579 | 36,171 | 57.8 | 45,967 | 73.5 |
| Otava | 90,191 | 66,022 | 73.2 | 71,093 | 78.8 |
| ASDI | 108,191 | 69,559 | 64.3 | 84,076 | 77.7 |
| Chemical Block | 121,063 | 81,295 | 67.2 | 88,737 | 73.3 |
| Sigma Aldrich | 165,860 | 80,055 | 48.3 | 112,079 | 67.6 |
| TimTec | 196,086 | 131,245 | 66.9 | 151,480 | 77.3 |
| Specs | 198,771 | 143,042 | 72.0 | 135,730 | 68.3 |
| AMRI | 253,405 | 231,132 | 91.2 | 134,512 | 53.1 |
| Alinda | 256,282 | 205,240 | 80.1 | 199,346 | 77.8 |
| Zelinsky Institute | 299,835 | 221,215 | 73.8 | 228,387 | 76.2 |
| Life Chemicals | 329,403 | 283,359 | 86.0 | 284,275 | 86.3 |
| ASINEX | 412,485 | 349,433 | 84.7 | 345,795 | 83.8 |
| InterBioScreen | 450,740 | 322,579 | 71.6 | 345,292 | 76.6 |
| PrincetonBio | 608,389 | 460,595 | 75.7 | 458,874 | 75.4 |
| ChemBridge | 634,160 | 513,978 | 81.0 | 538,243 | 84.9 |
| ChemDiv | 825,535 | 684,360 | 82.9 | 615,231 | 74.5 |
| Vitas M Lab | 924,849 | 667,441 | 72.2 | 687,876 | 74.4 |
| Abamachem | 1,433,640 | 1,241,030 | 86.6 | 1,387,656 | 96.8 |
| FCHGroup | 1,455,286 | 1,231,817 | 84.6 | 1,387,818 | 95.4 |
| BCHResearch | 1,499,151 | 1,259,359 | 84.0 | 1,435,948 | 95.8 |
| UORSY | 1,525,625 | 1,239,842 | 81.3 | 1,405,685 | 92.1 |
| Selena Chem | 1,537,890 | 1,294,915 | 84.2 | 1,472,515 | 95.7 |
| Enamine | 1,636,699 | 1,389,329 | 84.9 | 1,489,060 | 91.0 |
| NCI | 231,346 | 123,313 | 53.3 | 184,691 | 79.8 |
| **TOTAL (unique)** | **11,870,691** | **9,753,597** | **82.2** | **10,570,543** | **89.0** |

**Table S3** Continued

| Supplier | *Veber rule* | | *Number of aromatic rings ≤3* | | *Total filtered 'drug-like'* | |
|---|---|---|---|---|---|---|
| | **No. of compounds passed** | **%** | **No. of compounds passed** | **%** | **No. of compounds passed** | **%** |
| Focus Synthesis | 872 | 86.9 | 1,001 | 99.8 | 553 | 55.1 |
| Menai Organics | 3,564 | 99.1 | 3,422 | 95.2 | 1,558 | 43.3 |
| Peakdale | 15,111 | 99.3 | 12,390 | 81.4 | 10,441 | 68.6 |
| TosLab | 15,444 | 95.2 | 11,485 | 70.8 | 6,111 | 37.7 |
| MDPI | 16,108 | 87.7 | 16,494 | 89.8 | 7,414 | 40.4 |
| AllLab | 18,964 | 97.2 | 17,756 | 91.0 | 10,475 | 53.7 |
| Aronis | 22,111 | 96.4 | 21,750 | 94.9 | 11,253 | 49.1 |
| AnalytiCon | 19,629 | 81.0 | 23,056 | 95.1 | 15,526 | 64.1 |
| EMC | 23,306 | 87.7 | 22,778 | 85.7 | 15,078 | 56.7 |
| Intermed Ltd | 31,133 | 97.0 | 16,224 | 50.5 | 7,841 | 24.4 |
| Synthon-Lab | 44,352 | 95.9 | 40,684 | 87.9 | 13,970 | 30.2 |
| Key Organics | 51,447 | 98.9 | 47,370 | 91.0 | 27,958 | 53.7 |
| Maybridge | 59,435 | 97.6 | 56,364 | 92.5 | 32,495 | 53.3 |
| Nanosyn | 59,505 | 95.1 | 52,874 | 84.5 | 24,815 | 39.7 |
| Otava | 87,627 | 97.2 | 75,951 | 84.2 | 47,274 | 52.4 |
| ASDI | 104,030 | 96.2 | 85,142 | 78.7 | 47,770 | 44.2 |
| Chemical Block | 113,571 | 93.8 | 99,634 | 82.3 | 54,694 | 45.2 |
| Sigma Aldrich | 143,316 | 86.4 | 141,649 | 85.4 | 56,648 | 34.2 |
| TimTec | 185,115 | 94.4 | 168,150 | 85.8 | 95,390 | 48.6 |
| Specs | 189,888 | 95.5 | 161,868 | 81.4 | 88,646 | 44.6 |
| AMRI | 221,488 | 87.4 | 194,752 | 76.9 | 101,945 | 40.2 |
| Alinda | 246,049 | 96.0 | 228,815 | 89.3 | 149,419 | 58.3 |
| Zelinsky Institute | 286,910 | 95.7 | 258,289 | 86.1 | 154,568 | 51.6 |
| Life Chemicals | 323,072 | 98.1 | 247,428 | 75.1 | 189,063 | 57.4 |
| ASINEX | 394,796 | 95.7 | 341,138 | 82.7 | 254,466 | 61.7 |
| InterBioScreen | 432,016 | 95.8 | 335,449 | 74.4 | 195,692 | 43.4 |
| Princeton BioMol | 582,879 | 95.8 | 497,804 | 81.8 | 309,846 | 50.9 |
| ChemBridge | 615,981 | 97.1 | 565,845 | 89.2 | 403,155 | 63.6 |
| ChemDiv | 796,512 | 96.5 | 618,529 | 74.9 | 418,362 | 50.7 |
| Vitas M Lab | 886,029 | 95.8 | 735,217 | 79.5 | 431,310 | 46.6 |
| Abamachem | 1,399,610 | 97.6 | 1,322,101 | 92.2 | 1,093,327 | 76.3 |
| FCHGroup | 1,401,141 | 96.3 | 1,284,536 | 88.3 | 1,012,925 | 69.6 |
| BCHResearch | 1,448,875 | 96.6 | 1,344,990 | 89.7 | 1,060,458 | 70.7 |
| UORSY | 1,472,784 | 96.5 | 1,355,249 | 88.8 | 1,011,052 | 66.3 |
| Selena Chem | 1,485,499 | 96.6 | 1,374,872 | 89.4 | 1,085,639 | 70.6 |
| Enamine | 1,599,286 | 97.7 | 1,425,029 | 87.1 | 1,132,453 | 69.2 |
| NCI | 204,113 | 88.2 | 213,540 | 92.3 | 99,205 | 42.9 |
| **TOTAL (unique)** | 11,407,857 | 96.1 | 10,301,261 | 86.8 | 7,674,631 | 64.7 |

**Table S4** Analysis of the structural features of 'drug-like' compounds.

| Supplier | *Ph ≤1* | | *Fsp³ ≥0.3* | | *$f_{MF}$ filter*[a] | | *In total structural enrichment filtered* | | |
|---|---|---|---|---|---|---|---|---|---|
| | **No. of compounds passed** | **%**[b] | **No. of compounds passed** | **%**[b] | **No. of compounds passed** | **%**[b] | **No. of compounds passed** | **%**[b] | **%**[c] |
| Focus Synthesis | 517 | 93.5 | 258 | 46.7 | 415 | 75.0 | 158 | 28.6 | 15.8 |
| Menai Organics | 748 | 48.0 | 546 | 35.0 | 870 | 55.8 | 312 | 20.0 | 8.7 |
| Peakdale | 7,247 | 69.4 | 6,708 | 64.2 | 3,913 | 37.5 | 1,851 | 17.7 | 12.2 |
| TosLab | 3,843 | 62.9 | 2,954 | 48.3 | 3,289 | 53.8 | 1,321 | 21.6 | 8.1 |
| MDPI | 5,322 | 71.8 | 3,139 | 42.3 | 5,408 | 72.9 | 2,413 | 32.5 | 13.1 |
| AllLab | 5,226 | 49.9 | 3,905 | 37.3 | 7,305 | 69.7 | 2,181 | 20.8 | 11.2 |
| Aronis | 3,875 | 34.4 | 3,954 | 35.1 | 7,958 | 70.7 | 2,050 | 18.2 | 8.9 |
| AnalytiCon | 11,746 | 75.7 | 13,538 | 87.2 | 6,509 | 41.9 | 5,455 | 35.1 | 22.5 |
| EMC | 10,055 | 66.7 | 10,721 | 71.1 | 9,990 | 66.3 | 6,348 | 42.1 | 23.9 |
| Intermed Ltd | 4,237 | 54.0 | 1,197 | 15.3 | 2,841 | 36.2 | 731 | 9.3 | 2.3 |
| Synthon-Lab | 6,535 | 46.8 | 4,981 | 35.7 | 7,214 | 51.6 | 2,291 | 16.4 | 5.0 |
| Key Organics | 19,109 | 68.3 | 7,816 | 28.0 | 19,583 | 70.0 | 5,380 | 19.2 | 10.3 |
| Maybridge | 23,264 | 71.6 | 11,275 | 34.7 | 22,694 | 69.8 | 7,752 | 23.9 | 12.7 |
| Nanosyn | 13,214 | 53.3 | 8,491 | 34.2 | 16,735 | 67.4 | 5,422 | 21.8 | 8.7 |
| Otava | 24,502 | 51.8 | 19,919 | 42.1 | 30,252 | 64.0 | 10,704 | 22.6 | 11.9 |
| ASDI | 26,752 | 56.0 | 16,408 | 34.3 | 31,647 | 66.2 | 10,844 | 22.7 | 10.0 |
| Chemical Block | 32,258 | 59.0 | 22,951 | 42.0 | 34,410 | 62.9 | 13,292 | 24.3 | 11.0 |
| Sigma Aldrich | 38,296 | 67.6 | 24,305 | 42.9 | 42,705 | 75.4 | 17,127 | 30.2 | 10.3 |
| TimTec | 58,618 | 61.5 | 38,914 | 40.8 | 63,551 | 66.6 | 23,842 | 25.0 | 12.2 |
| Specs | 44,150 | 49.8 | 33,389 | 37.7 | 59,677 | 67.3 | 19,789 | 22.3 | 10.0 |
| AMRI | 48,255 | 47.3 | 73,933 | 72.5 | 49,428 | 48.5 | 26,499 | 26.0 | 10.5 |
| Alinda | 73,842 | 49.4 | 66,016 | 44.2 | 94,765 | 63.4 | 34,722 | 23.2 | 13.5 |
| Zelinsky Institute | 95,196 | 61.6 | 74,820 | 48.4 | 100,616 | 65.1 | 44,745 | 28.9 | 14.9 |
| Life Chemicals | 99,179 | 52.5 | 85,727 | 45.3 | 92,733 | 49.0 | 35,620 | 18.8 | 10.8 |
| ASINEX | 166,037 | 65.2 | 158,139 | 62.1 | 121,405 | 47.7 | 60,057 | 23.6 | 14.6 |
| InterBioScreen | 112,527 | 57.5 | 95,747 | 48.9 | 114,049 | 58.3 | 51,200 | 26.2 | 11.4 |
| Princeton Bio | 159,633 | 51.5 | 137,974 | 44.5 | 180,362 | 58.2 | 68,822 | 22.2 | 11.3 |
| ChemBridge | 222,053 | 55.1 | 227,677 | 56.5 | 216,889 | 53.8 | 94,169 | 23.4 | 14.8 |
| ChemDiv | 202,450 | 48.4 | 221,289 | 52.9 | 210,601 | 50.3 | 82,820 | 19.8 | 10.0 |
| Vitas M Lab | 219,804 | 51.0 | 190,419 | 44.1 | 256,736 | 59.5 | 96,369 | 22.3 | 10.4 |
| Abamachem | 618,708 | 56.6 | 683,344 | 62.5 | 606,873 | 55.5 | 299,899 | 27.4 | 20.9 |
| FCHGroup | 460,064 | 45.4 | 635,652 | 62.8 | 489,435 | 48.3 | 196,792 | 19.4 | 13.5 |
| BCHResearch | 519,540 | 49.0 | 648,259 | 61.1 | 550,981 | 52.0 | 235,627 | 22.2 | 15.7 |
| UORSY | 452,374 | 44.7 | 533,776 | 52.8 | 576,116 | 57.0 | 208,335 | 20.6 | 13.7 |
| Selena Chem | 513,818 | 47.3 | 662,544 | 61.0 | 551,136 | 50.8 | 226,462 | 20.9 | 14.7 |
| Enamine | 641,605 | 56.7 | 608,702 | 53.8 | 591,148 | 52.2 | 256,455 | 22.6 | 15.7 |
| NCI | 71,349 | 71.9 | 52,059 | 52.5 | 75,045 | 75.6 | 35,740 | 36.0 | 15.4 |
| **TOTAL (unique)** | 3,965,527 | 51.7 | 4,512,905 | 58.8 | 4,046,545 | 52.7 | 1,747,796 | 22.8 | 14.7 |

[a] threshold values for $f_{MF}$ filtering: <0.95 for 1 ring system (RS), <0.85 for 2 RS, and <0.75 for 3+ RS.
[b] calculated in regard to the 'drug-like' compounds
[c] calculated in regard to the total size of compound libraries

**Table S5** Analysis of the attrition rate of scaffolds and topological patterns after structural enrichment in the libraries of individual suppliers

| Supplier | Scaffolds, total | Topological patterns, total | Scaffolds, 'drug-like' | Topological patterns, 'drug-like' | % structurally enriched to total | | % structurally enriched to 'drug-like' | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Scaffolds | Topological patterns | Scaffolds | Topological patterns |
| Focus Synthesis | 384 | 186 | 252 | 131 | 26.04 | 30.65 | 39.68 | 43.51 |
| Menai Organics | 1,266 | 510 | 602 | 256 | 10.19 | 12.75 | 21.43 | 25.39 |
| Peakdale | 5,280 | 1,967 | 3,565 | 1,236 | 11.31 | 11.79 | 16.75 | 18.77 |
| TosLab | 6,006 | 2,808 | 2,493 | 1,038 | 10.37 | 8.12 | 24.99 | 21.97 |
| MDPI | 5,293 | 1,965 | 2,040 | 581 | 12.60 | 8.50 | 32.70 | 28.74 |
| AllLab | 3,313 | 1,013 | 1,820 | 552 | 12.89 | 12.73 | 23.46 | 23.37 |
| Aronis | 3,633 | 1,238 | 1,427 | 475 | 7.07 | 6.62 | 18.01 | 17.26 |
| AnalytiCon | 8,145 | 3,369 | 4,843 | 1,803 | 16.61 | 15.20 | 27.94 | 28.40 |
| EMC | 3,813 | 1,810 | 2,296 | 1,024 | 15.11 | 12.82 | 25.09 | 22.66 |
| Intermed Ltd | 2,660 | 1,072 | 615 | 227 | 7.26 | 6.90 | 31.38 | 32.60 |
| Synthon-Lab | 6,170 | 2,121 | 1,646 | 579 | 6.09 | 5.66 | 22.84 | 20.73 |
| Key Organics | 12,760 | 2,963 | 6,913 | 1,342 | 13.31 | 12.01 | 24.56 | 26.53 |
| Maybridge | 18,607 | 4,204 | 9,896 | 1,931 | 12.83 | 10.82 | 24.13 | 23.56 |
| Nanosyn | 15,633 | 5,152 | 6,089 | 1,735 | 8.78 | 6.29 | 22.53 | 18.67 |
| Otava | 18,274 | 5,736 | 9,117 | 2,408 | 10.63 | 8.32 | 21.30 | 19.81 |
| ASDI | 22,115 | 6,191 | 10,593 | 2,418 | 11.35 | 8.79 | 23.70 | 22.50 |
| Chemical Block | 35,450 | 11,375 | 16,715 | 4,256 | 11.78 | 8.15 | 24.98 | 21.78 |
| Sigma Aldrich | 29,864 | 9,885 | 10,009 | 2,134 | 12.16 | 7.26 | 36.27 | 33.65 |
| TimTec | 50,935 | 14,424 | 23,869 | 5,304 | 11.66 | 7.89 | 24.88 | 21.46 |
| Specs | 44,774 | 14,107 | 16,505 | 3,913 | 8.17 | 5.94 | 22.15 | 21.42 |
| AMRI | 27,686 | 12,952 | 13,693 | 6,015 | 9.40 | 7.71 | 19.00 | 16.59 |
| Alinda | 40,420 | 10,432 | 20,742 | 4,532 | 10.32 | 7.47 | 20.11 | 17.19 |
| Zelinsky Institute | 50,802 | 13,867 | 21,731 | 4,827 | 10.96 | 8.16 | 25.61 | 23.45 |
| Life Chemicals | 60,227 | 15,775 | 34,364 | 7,702 | 10.41 | 7.39 | 18.24 | 15.14 |
| ASINEX | 99,551 | 24,119 | 64,225 | 13,985 | 13.12 | 9.76 | 20.33 | 16.84 |
| InterBioScreen | 86,916 | 22,741 | 40,035 | 9,325 | 11.63 | 8.67 | 25.26 | 21.15 |
| PrincetonBio | 95,034 | 23,572 | 43,542 | 8,774 | 10.08 | 6.69 | 22.01 | 17.97 |
| ChemBridge | 146,146 | 33,285 | 92,653 | 19,279 | 11.86 | 7.63 | 18.71 | 13.16 |
| ChemDiv | 131,630 | 34,990 | 71,404 | 16,807 | 11.26 | 7.52 | 20.75 | 15.65 |
| Vitas M Lab | 145,993 | 34,567 | 64,114 | 12,413 | 9.65 | 6.58 | 21.97 | 18.31 |
| Abamachem | 402,579 | 45,996 | 275,646 | 31,005 | 12.69 | 8.66 | 18.53 | 12.85 |
| FCHGroup | 460,864 | 55,830 | 292,867 | 36,284 | 10.41 | 8.08 | 16.38 | 12.44 |
| BCHResearch | 441,042 | 53,696 | 283,495 | 34,916 | 11.07 | 8.13 | 17.22 | 12.50 |
| UORSY | 311,468 | 48,351 | 187,833 | 27,303 | 10.94 | 7.68 | 18.15 | 13.59 |
| Selena Chem | 454,259 | 54,730 | 291,590 | 35,826 | 10.82 | 8.09 | 16.86 | 12.35 |
| Enamine | 457,510 | 62,545 | 283,553 | 33,881 | 11.87 | 7.13 | 19.15 | 13.16 |
| NCI | 60,962 | 17,104 | 25,369 | 5,111 | 15.58 | 9.48 | 37.43 | 31.72 |
| **Combined database** | 1,857,029 | 196,439 | 1,116,697 | 105,127 | 10.03 | 6.58 | 16.67 | 12.29 |

**Table S6** Diversity analysis of the structurally enriched compounds selected from the commercial libraries

| Supplier | Scaffolds | Self-dissimilarity of scaffolds | Average compounds per scaffold | Topological patterns | Compounds clustered by topological patterns | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | singletons | small (2-10) | medium (11-100) | large (>100) | Clustering coverage, % | Average cluster size |
| Focus Synthesis | 100 | 0.8404 | 1.58 | 57 | 42 | 12 | 3 | 0 | 73.41 | 2.77 |
| Menai Organics | 129 | 0.8124 | 2.42 | 65 | 26 | 8 | 31 | 0 | 91.66 | 4.80 |
| Peakdale | 597 | 0.8148 | 3.10 | 232 | 91 | 98 | 40 | 3 | 95.08 | 7.98 |
| TosLab | 623 | 0.8524 | 2.12 | 228 | 97 | 109 | 22 | 0 | 92.65 | 5.79 |
| MDPI | 667 | 0.8549 | 3.62 | 167 | 78 | 68 | 16 | 5 | 96.76 | 14.45 |
| AllLab | 427 | 0.8395 | 5.11 | 129 | 34 | 63 | 26 | 6 | 98.44 | 16.91 |
| Aronis | 257 | 0.8220 | 7.98 | 82 | 31 | 29 | 19 | 3 | 98.48 | 25.00 |
| AnalytiCon | 1,353 | 0.8329 | 4.03 | 512 | 147 | 252 | 107 | 6 | 97.3 | 10.65 |
| EMC | 576 | 0.8289 | 11.02 | 232 | 68 | 86 | 68 | 10 | 98.92 | 27.36 |
| Intermed Ltd | 193 | 0.8477 | 3.79 | 74 | 31 | 30 | 12 | 1 | 95.75 | 9.88 |
| Synthon-Lab | 376 | 0.8389 | 6.09 | 120 | 28 | 63 | 23 | 6 | 98.77 | 19.09 |
| Key Organics | 1,698 | 0.8372 | 3.17 | 356 | 145 | 164 | 36 | 11 | 97.3 | 15.11 |
| Maybridge | 2,388 | 0.8440 | 3.25 | 455 | 196 | 189 | 58 | 12 | 97.47 | 17.04 |
| Nanosyn | 1,372 | 0.8500 | 3.95 | 324 | 138 | 132 | 43 | 11 | 97.45 | 16.73 |
| Otava | 1,942 | 0.8379 | 5.51 | 477 | 181 | 197 | 82 | 17 | 98.3 | 22.44 |
| ASDI | 2,511 | 0.8458 | 4.32 | 544 | 259 | 202 | 68 | 15 | 97.61 | 19.93 |
| Chemical Block | 4,175 | 0.8432 | 3.18 | 927 | 389 | 392 | 125 | 21 | 97.07 | 14.34 |
| Sigma Aldrich | 3,630 | 0.8515 | 4.72 | 718 | 339 | 278 | 84 | 17 | 98.02 | 23.85 |
| TimTec | 5,939 | 0.8460 | 4.01 | 1,138 | 481 | 466 | 159 | 32 | 97.98 | 20.95 |
| Specs | 3,656 | 0.8417 | 5.41 | 838 | 347 | 345 | 122 | 24 | 98.24 | 23.61 |
| AMRI | 2,602 | 0.8182 | 10.18 | 998 | 282 | 432 | 230 | 54 | 98.93 | 26.55 |
| Alinda | 4,171 | 0.8339 | 8.32 | 779 | 282 | 300 | 160 | 37 | 99.18 | 44.57 |
| Zelinsky Institute | 5,566 | 0.8407 | 8.04 | 1,132 | 378 | 453 | 253 | 48 | 99.15 | 39.53 |
| Life Chemicals | 6,267 | 0.8357 | 5.68 | 1,166 | 347 | 492 | 252 | 75 | 99.02 | 30.55 |
| ASINEX | 13,058 | 0.8372 | 4.60 | 2,355 | 792 | 1,014 | 452 | 97 | 98.68 | 25.50 |
| InterBioScreen | 10,112 | 0.8463 | 5.06 | 1,972 | 723 | 811 | 356 | 82 | 98.58 | 25.96 |
| PrincetonBio | 9,583 | 0.8418 | 7.18 | 1,577 | 539 | 632 | 325 | 81 | 99.21 | 43.64 |
| ChemBridge | 17,332 | 0.8313 | 5.43 | 2,538 | 971 | 1,035 | 426 | 106 | 98.96 | 37.10 |
| ChemDiv | 14,818 | 0.8339 | 5.59 | 2,630 | 875 | 1,154 | 473 | 128 | 98.94 | 31.49 |
| Vitas M Lab | 14,084 | 0.8419 | 6.84 | 2,273 | 788 | 951 | 422 | 112 | 99.18 | 42.40 |
| Abamachem | 51,089 | 0.8182 | 5.87 | 3,985 | 1,410 | 1,600 | 718 | 257 | 99.52 | 75.26 |
| FCHGroup | 47,972 | 0.8172 | 4.10 | 4,512 | 1,572 | 1,899 | 778 | 263 | 99.2 | 43.62 |
| BCHResearch | 48,809 | 0.8183 | 4.83 | 4,365 | 1,543 | 1,782 | 785 | 255 | 99.34 | 53.98 |
| UORSY | 34,087 | 0.8242 | 6.11 | 3,711 | 1,352 | 1,472 | 661 | 226 | 99.35 | 56.14 |
| Selena Chem | 49,173 | 0.8178 | 4.61 | 4,426 | 1,533 | 1,841 | 788 | 264 | 99.32 | 51.17 |
| Enamine | 54,294 | 0.8299 | 4.72 | 4,459 | 1,666 | 1,784 | 758 | 251 | 99.35 | 57.51 |
| NCI | 9,495 | 0.8520 | 3.76 | 1,621 | 871 | 562 | 153 | 35 | 97.56 | 22.05 |
| **Combined database** | 186,177 | 0.8303 | 9.39 | 12,921 | 4,310 | 5,098 | 2,575 | 935 | 99.75 | 135.27 |

**Fig. 1** Comparison of the overlap of compounds from all suppliers in chemical space (scale corresponds to the percentage of overlapped compounds)

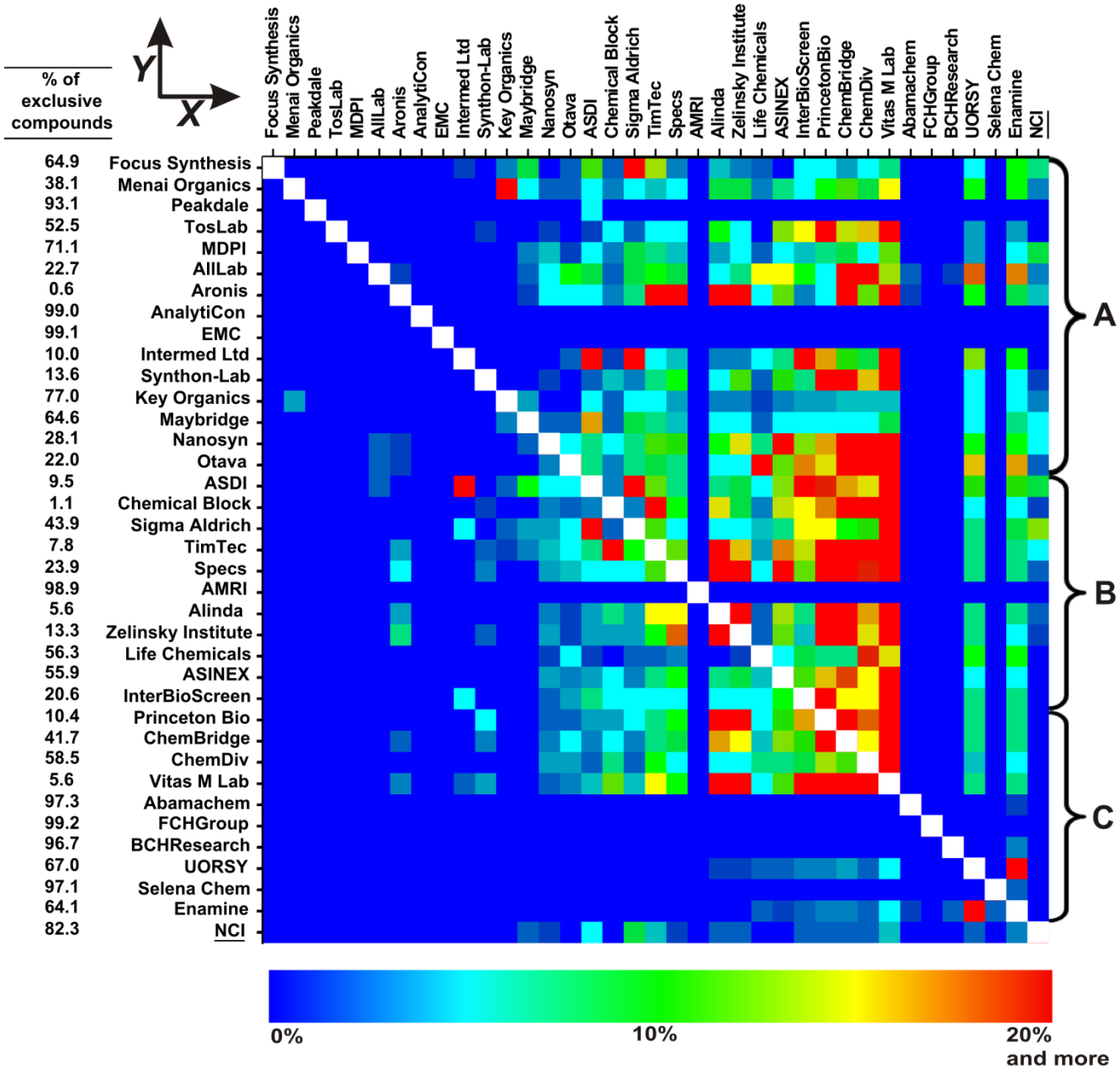


**Fig. 2** Summary of functional groups/substructures occurrences in the combined dataset of 11,870,691 unique structures. See the following pages S14 and S15 for the zoomed bar-graphs of the 'frequent' and 'rare' filters - Figures 2a and 2b, respectively.

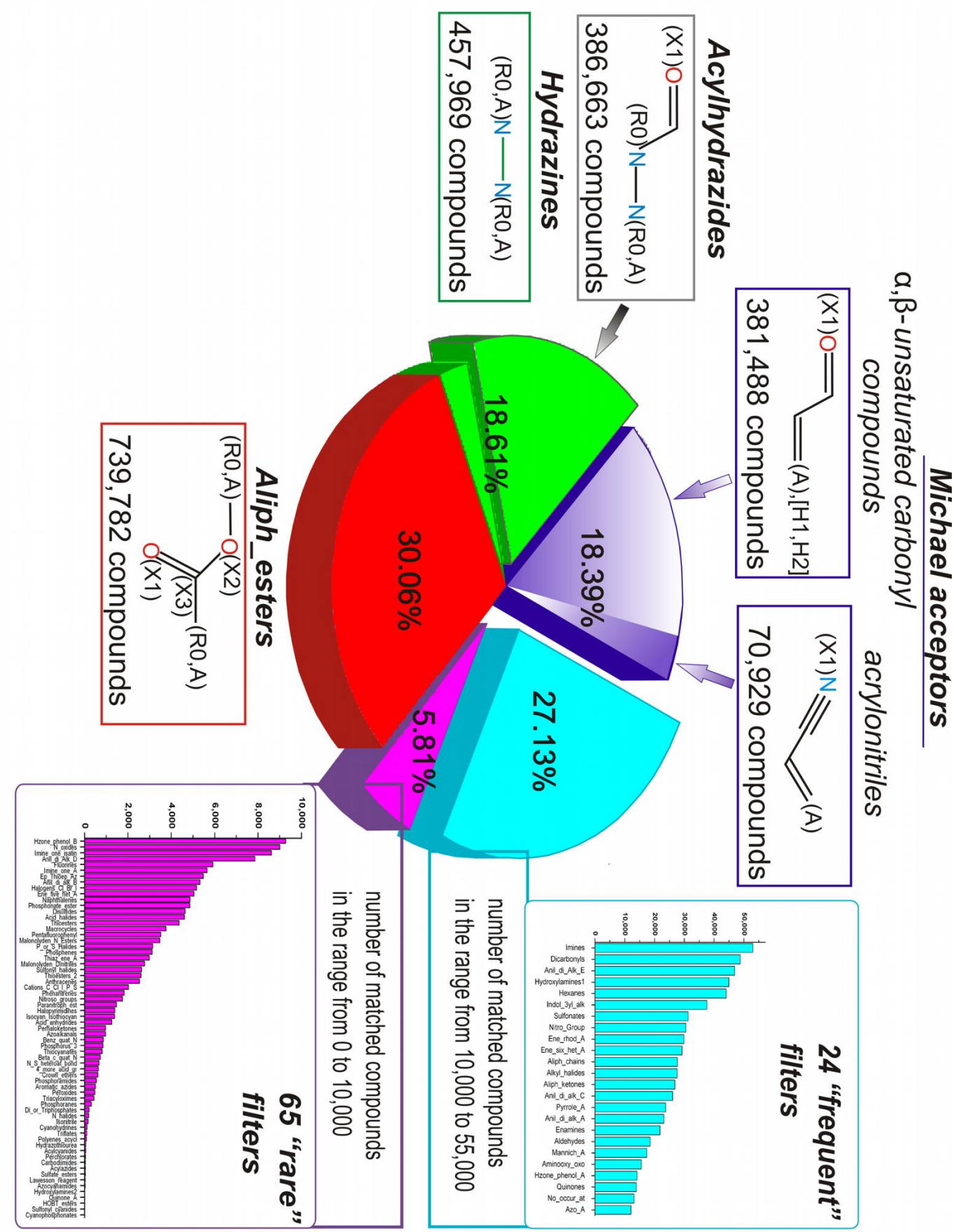


**Fig. 2a** Functional group/substructure occurrences across the compound dataset with number of matched compounds in the range from 10,000 to 55,000 (‘frequent’ filters).

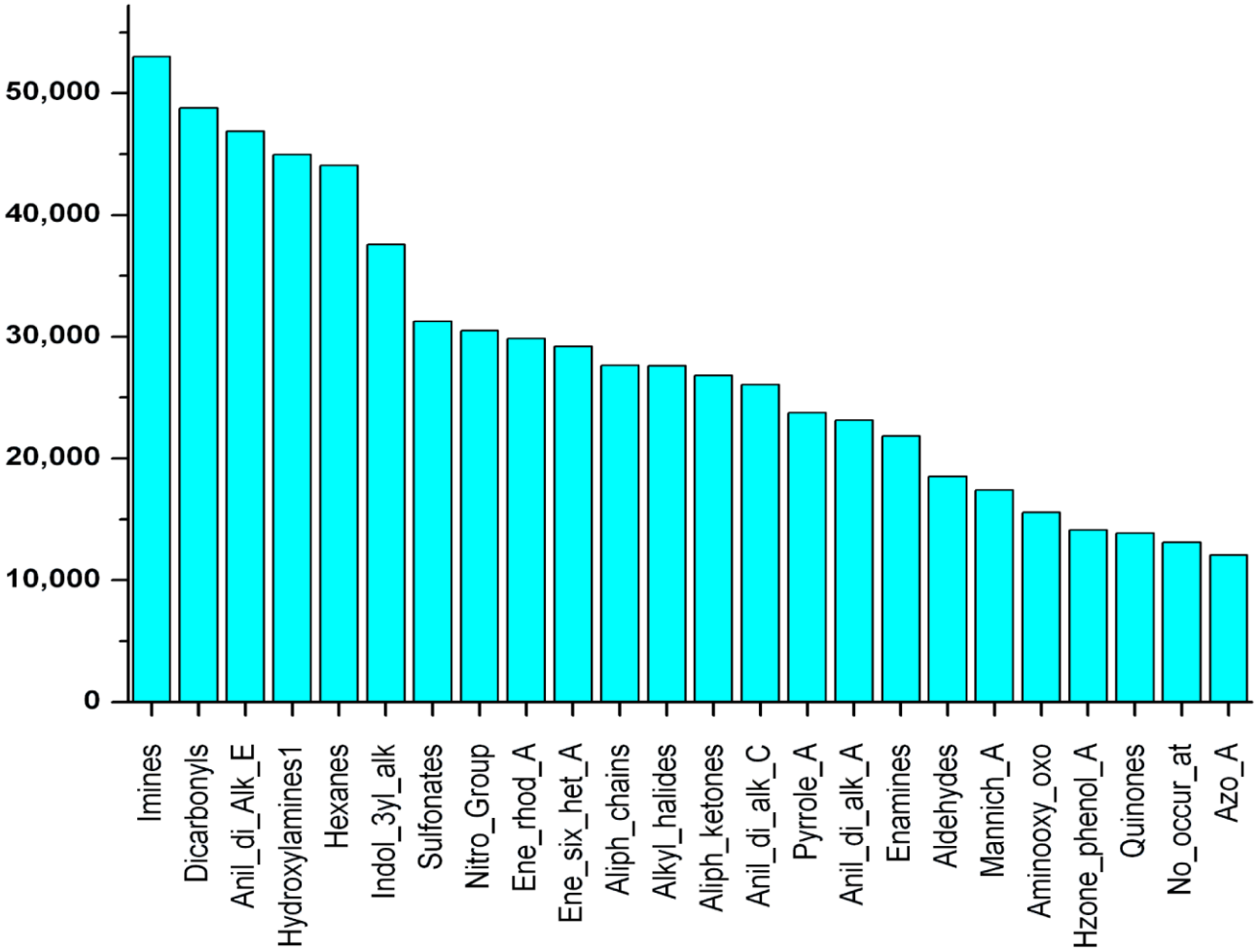


**Fig. 2b** Functional group/substructure occurrences across the compound dataset with number of matched compounds in the range from 0 to 10,000 (‘rare’ filters).

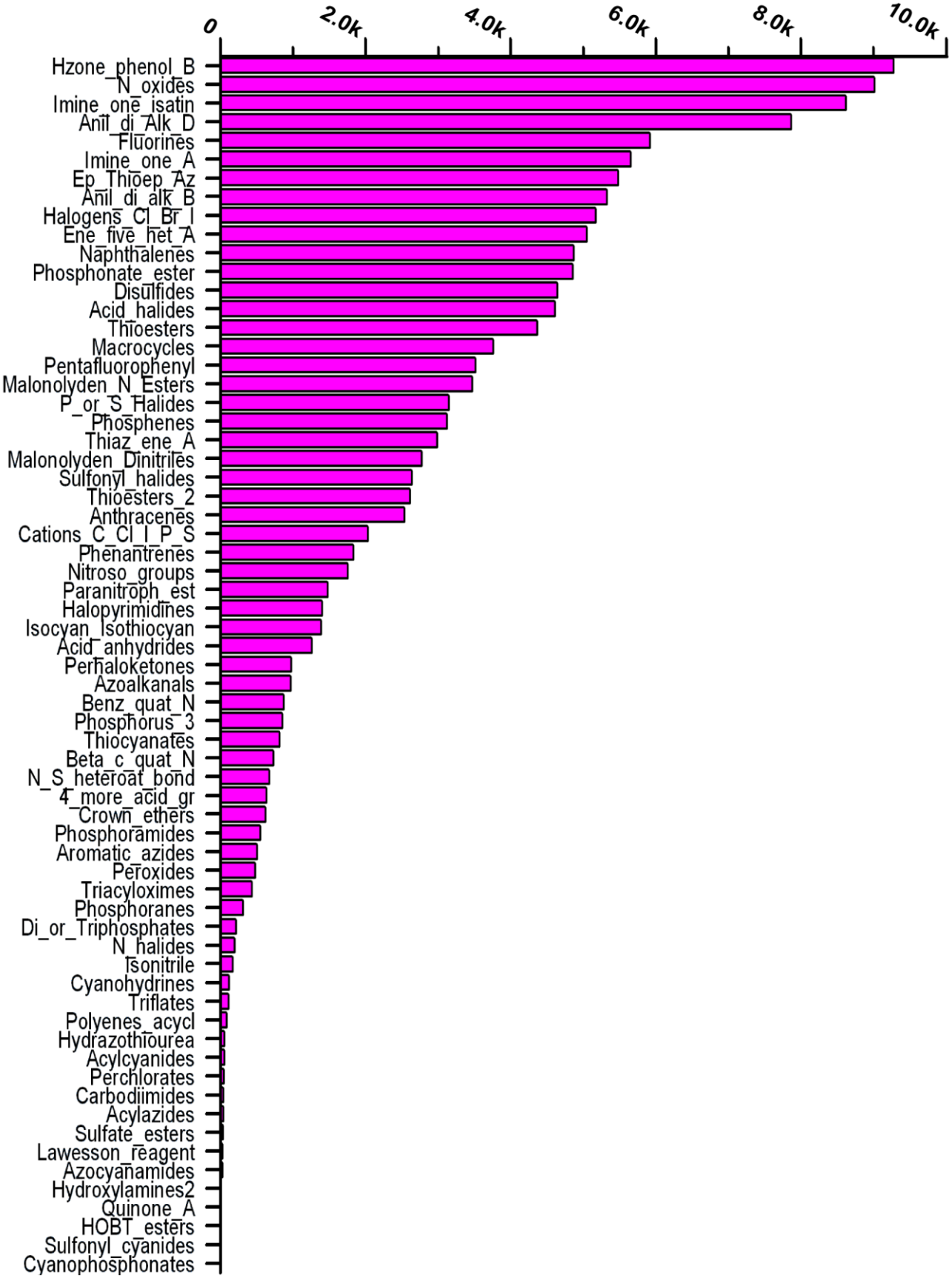


**Fig. S1**: The results of substructure filtering for libraries of individual suppliers.

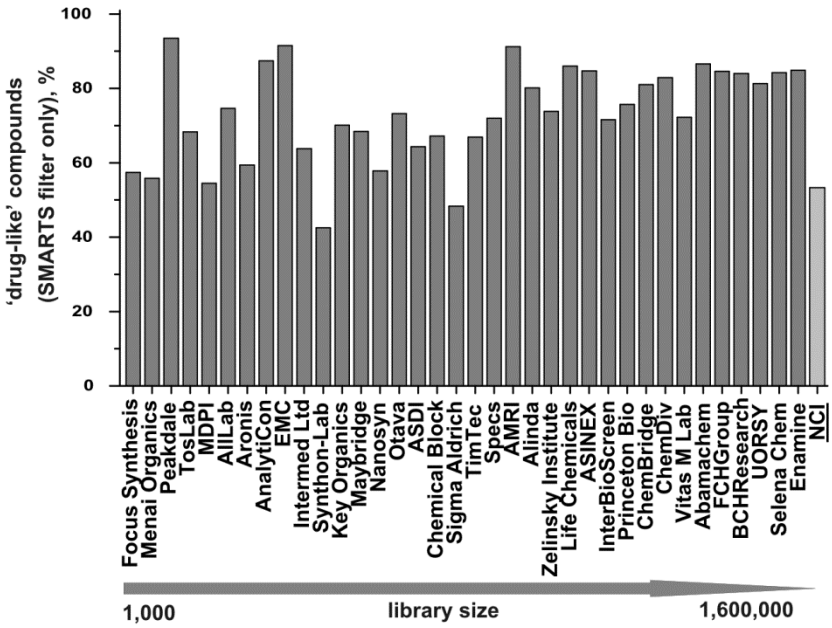



**Fig. S2** Distribution of the calculated physicochemical properties in the whole dataset.

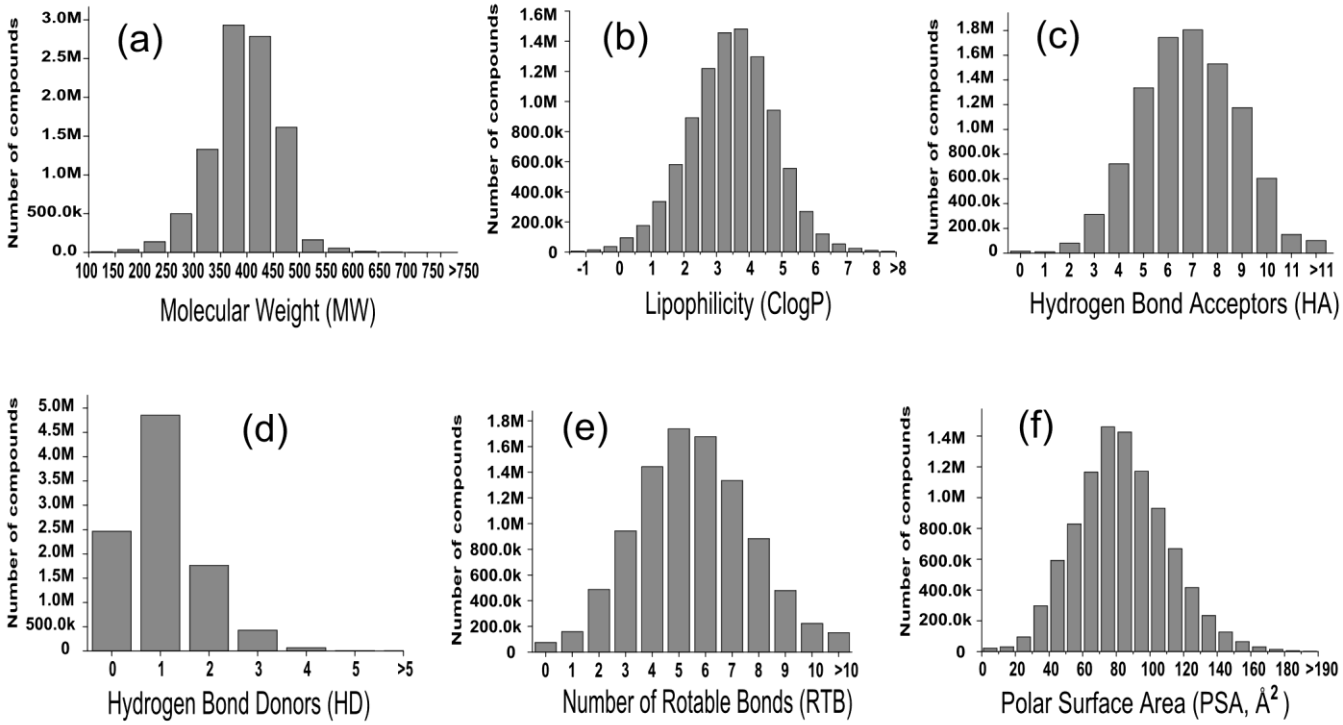


**Fig. 8** Proportion of scaffolds and topological patterns of structurally enriched compounds to the scaffolds and topological patterns of 'drug-like' compounds in the libraries of individual suppliers

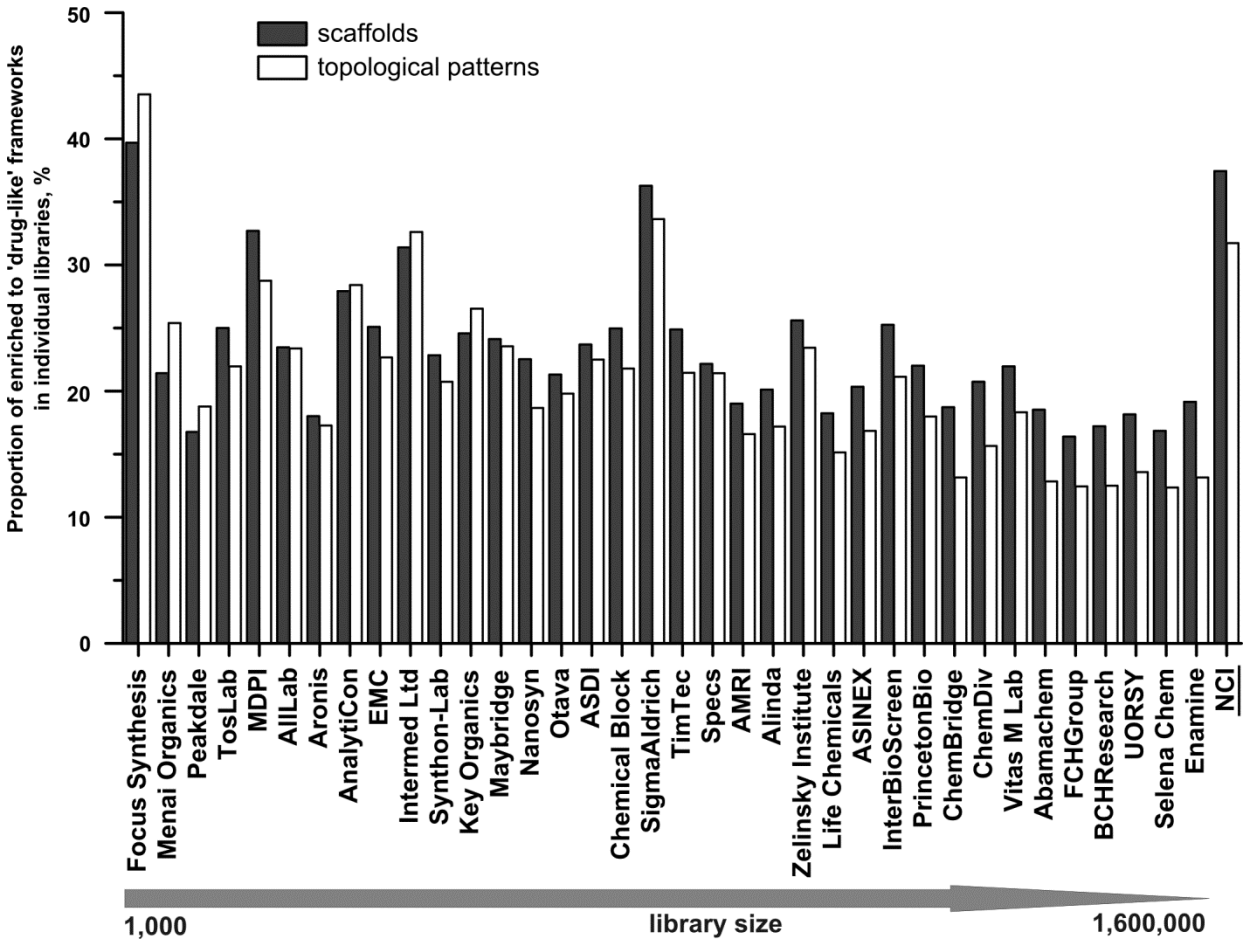


**Fig. 9** Distribution of cluster sizes and percentage of compounds in the corresponding clusters of the selection of structurally enriched compounds from the commercial libraries

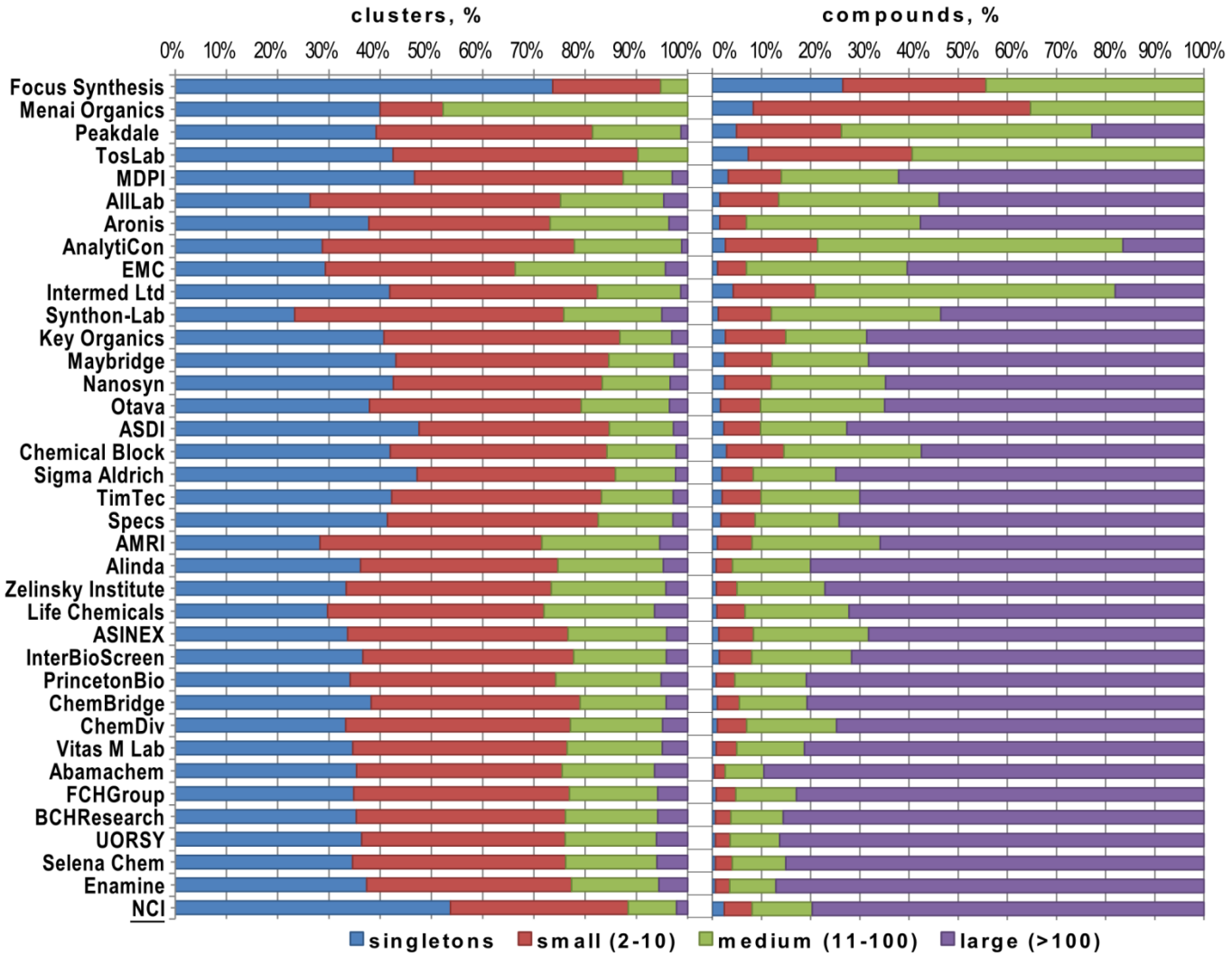


**Fig. 10**(b) Representation of the combined database of structurally enriched compounds by each library (by scaffolds and topological patterns)

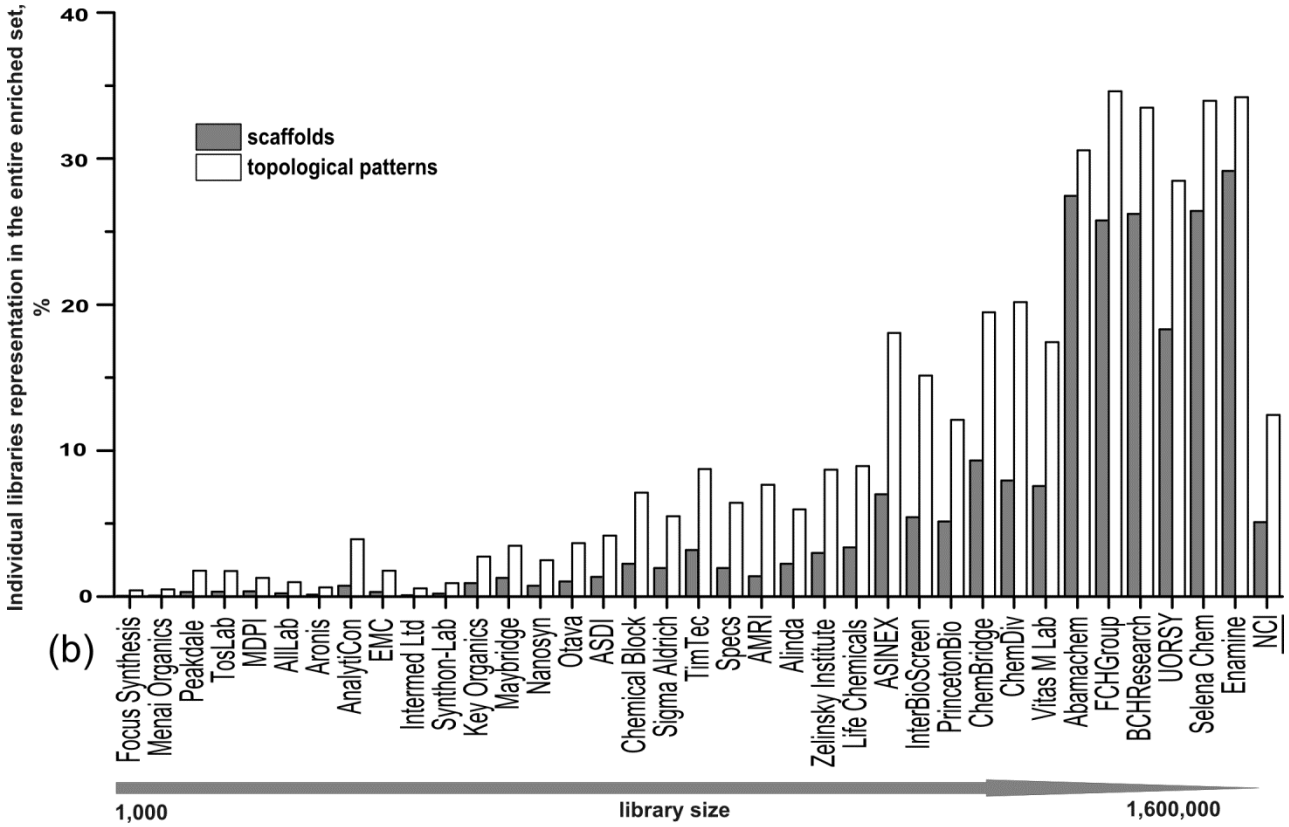


**Fig. 11** Comparison of the overlap of scaffolds that represent compounds remained after structural enrichment (scale corresponds to the percentage of overlapped scaffolds)

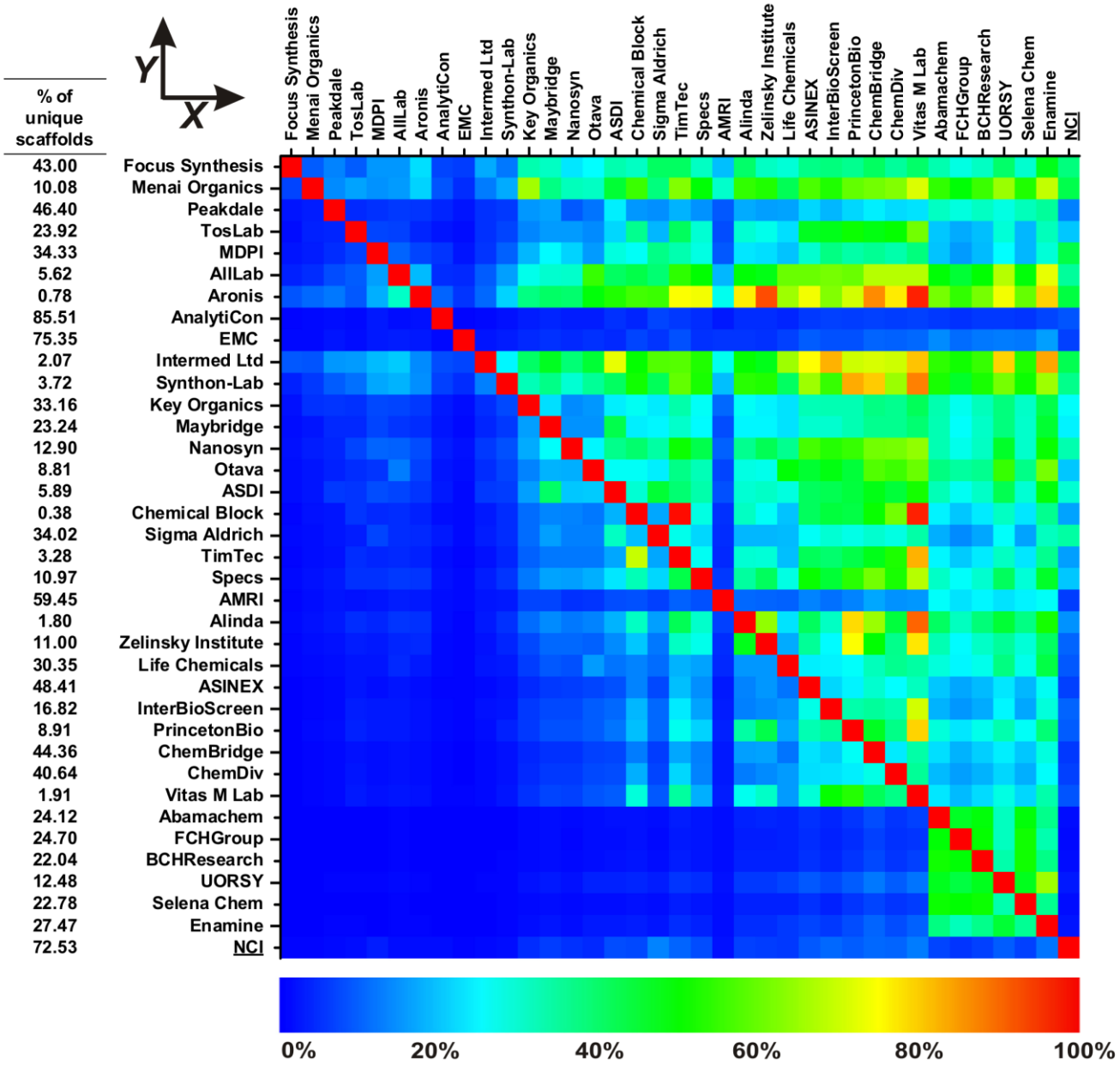
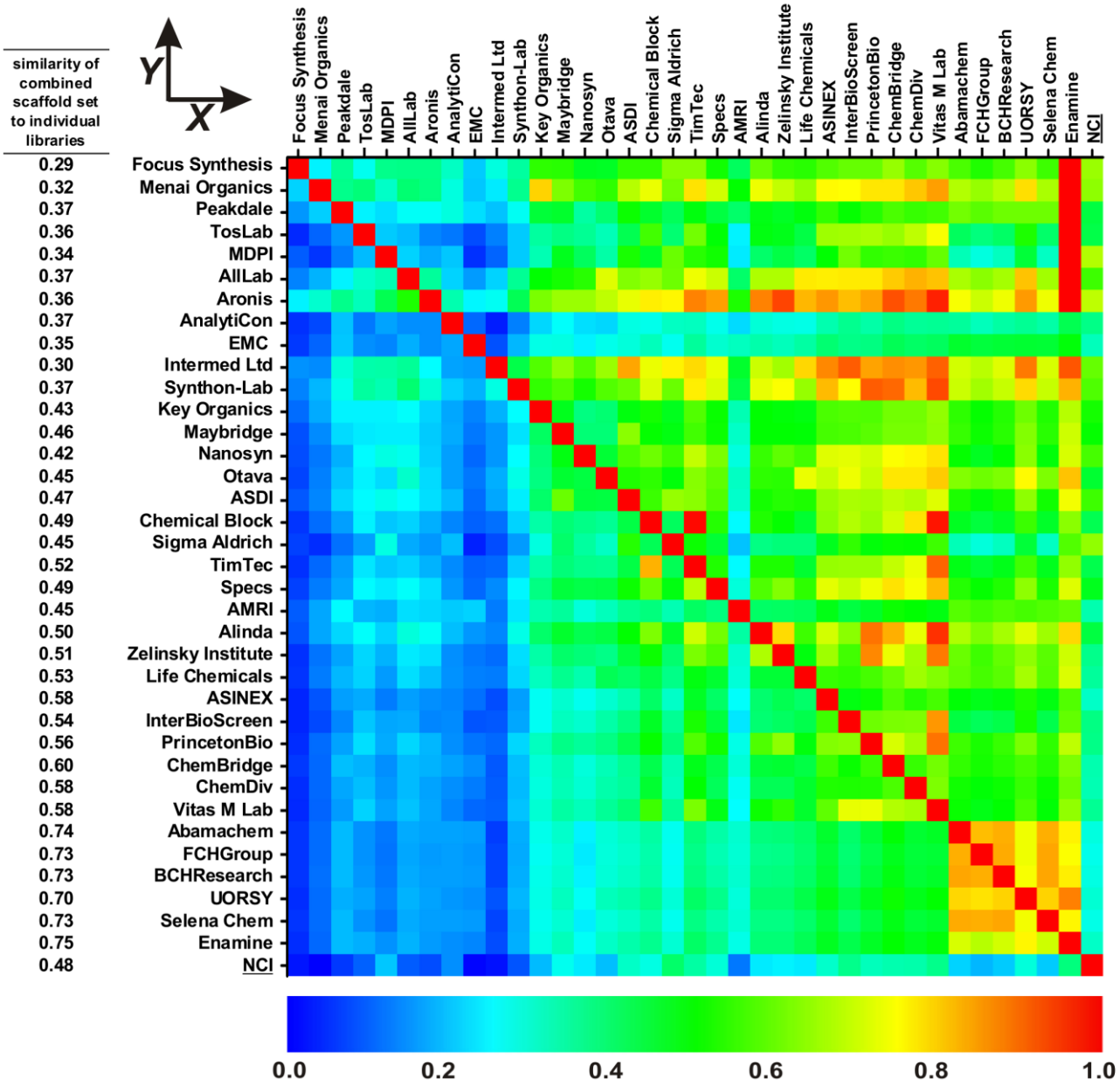


**Fig. 12** Similarity analysis of scaffolds that represent compounds remained after structural enrichment



## References


1. InstantJChem, 5.5.1.0, ChemAxon, Budapest, Hungary, http://www.chemaxon.com, 2011.
2. MySQL, Community Server 5.5, Oracle, https://www.mysql.com, 2011.
3. JChem, 5.5.1.0, ChemAxon, Budapest, Hungary, http://www.chemaxon.com, 2011.
4. S. V. Trepalin and A. V. Yarkov, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 100-107.
5. Daylight, Theory Manual, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, USA, http://www.daylight.com/dayhtml/doc/theory/index.html, 2011.
6. Marvin, 5.5.1.0, ChemAxon, Budapest, Hungary, http://www.chemaxon.com, 2011.
7. F. Lovering, J. Bikker and C. Humblet, *J. Med. Chem.*, 2009, **52**, 6752-6756.
8. Y. Yang, H. Chen, I. Nilsson, S. Muresan and O. Engkvist, *J. Med. Chem.*, 2010, **53**, 7709-7714.
9. W. R. Pitt, D. M. Parry, B. G. Perry and C. R. Groom, *J. Med. Chem.*, 2009, **52**, 2952-2963.
10. T. J. Ritchie and S. J. F. Macdonald, *Drug Discovery Today*, 2009, **14**, 1011-1020.
11. T. J. Ritchie, S. J. F. Macdonald, R. J. Young and S. D. Pickett, *Drug Discovery Today*, 2011, **16**, 164-171.
12. D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742-754.
13. J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea and B. K. Shoichet, *J. Chem. Inf. Model.*, 2008, **48**, 755-765.
14. G. Papadatos, A. W. J. Cooper, V. Kadirkamanathan, S. J. F. Macdonald, I. M. McLay, S. D. Pickett, J. M. Pritchard, P. Willett and V. J. Gillet, *J. Chem. Inf. Model.*, 2008, **49**, 195-208.
15. M. Vogt, D. Stumpfe, H. Geppert and J. Bajorath, *J. Med. Chem.*, 2010, **53**, 5707-5715.
16. W. P. Walters and M. A. Murcko, *Adv. Drug Delivery Rev.*, 2002, **54**, 255-271.