Supporting Information

Biased Complement Diversity Selection for Effective Exploration of Chemical Space in Hit Finding Campaigns

Johanna M. Jansen,* Gianfranco De Pascale, Susan Fong, Mika Lindvall, Heinz E. Moser, Keith Pfister, Bob Warne, Charles Wartchow

* Corresponding author: johanna.jansen@novartis.com

Contents

Note: reference numbers refer to the reference list in the paper

Workflow documentation

The three major steps in the workflow are summarized in Figure S1 and include creation and annotation of a compound library, a quality assessment of every cluster of compounds based on the annotations, and a biased selection of representative compounds from each cluster. Creation of the compound library starts with a list of unique chemical structures from which the subset should be selected; a unique identifier and a structure description using SMILES are expected. The chemical space annotation is typically done by clustering of the 2D chemical fingerprints of the compounds, resulting in a cluster-number associated with each compound. The quality attributes and exclusions are the annotations that are project specific. Exclusions are used when there are attributes that are known to be incompatible with the concept of a quality hit, *e.g.* matches to substructures known to interfere with the assay, knowledge of the compound being a frequent hitter. Compounds with such attributes should be excluded from the workflow at the start. Remaining compounds are assigned to up to four separate classes, labeled A through D. The highest quality class is A, followed by B, C, and D. Attributes used in this classification often include calculated physical chemical properties, and substructure matching such that compounds with the most desirable attributes are assigned to the higher quality class. Projects that want to use a complement design option around a core-set should designate all the compounds in the core-set as class-A. The workflow will then select compounds (from classes B, C, and D) to complement the chemical space that is already covered by the class-A compounds in the core-set.

Based on the fully annotated dataset with a minimum of four columns (identifier, SMILES, cluster, and class), the workflow loops over every cluster and assesses the quality of the cluster. All singletons are passed into the final selection. If a cluster contains only class-D compounds, a single best representative is selected. The user can include a score (as an optional fifth column) to pick the single best compound in these all-D clusters, otherwise the workflow picks the first compound in the cluster. Two variables that are described in the KNIME workflow annotation ("selectD" and "minimizeD") govern this behavior. For clusters that have more than one compound with a class A, B, or C label, a biased selection occurs across three tiers in the third part of the workflow.

The user determines a coverage number and coverage type for the workflow, which governs how many compounds are selected from each of the clusters (variables "coverage" and "coverageType"). The coverage is always related to the number of compounds left in the cluster after class-D compounds have been removed; this is called the "group-size". There are two types of coverage available: one is relating the coverage number as a fraction of the group-size (coverageType = "P") and the other is relating the coverage number as a fraction of the square root of the group-size (coverageType = "S"). The latter is particularly useful when selecting a small set from a large collection where there are very large clusters that would dominate if the selection were driven by the group-size. As an example, if the user sets the coverage-type to P and coverage number to 0.1, then 10% of each group will be selected. If coverage-type is S and coverage number is 0.1, then 10% of the square root of the group-size will be selected.

If the desired number of molecules from a cluster (based on coverage number and coverage type) can be covered with available class-A molecules in the cluster, the protocol will do a diversity pick from that subset. If there are not enough class-A compounds, the protocol will take whatever class-A molecules are available (if any) and complement with class-B. If there is still not enough coverage, the protocol will pick from class-C. The actual picking is done using the RDKit Diversity Picker node in KNIME: This node picks diverse rows from an input table based on Tanimoto distance between fingerprints (Morgan fingerprints, Radius 2, 2048 bit length); the picking is done using the MaxMin algorithm.[15] The node has a complement option, which is used when picking from class-B to complement around available class-A compounds and when picking from class-C to complement around available class-A plus class-B compounds. This option has the effect of seeding the diversity pick. If a complement design is requested, all the compounds in class-A get selected automatically (they are the core-set) and the design is filled out with the remaining classes.

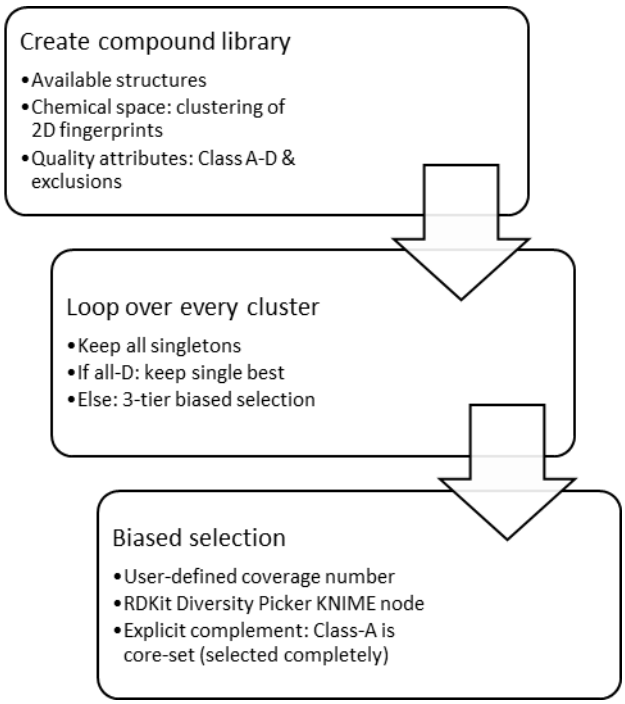

Figure S1. The components of the BCD workflow

Analysis of the relationship between clogD and cytotoxicity for compounds that inhibit the growth of Gram-negative bacteria

We collected a set of 985 validated inhibitors of Gram-negative (GN) bacteria (having an EC_{50} < 20 μM in at least one GN strain) that had been assessed for cytotoxic activity against two mammalian cell-lines (HepG2 and K562). These compounds were collected from our historical knowledgebase spanning many years and many projects. Compounds were labeled as not cytotoxic if cytotoxicity EC_{50} / GN EC_{50} ≥ 5 OR if cytotoxicity EC_{50} has a “>” qualifier. This condition had to be met for all bacterial strains that were tested in order to be labeled as not cytotoxic; if a compound was cytotoxic against one but not both mammalian cell-lines, it was still labeled as cytotoxic. In this data set, 87% of all compounds with clogD > 3 are cytotoxic (319 out of 368), and only 31% of the compounds in the full set are free of cytotoxicity (310 out of 985).

Case study details - Bacteriology

Exclusions
Substructure “Out”
Anything with properties not in this range: 250 ≤ MW ≤ 500; ROT ≤ 10; PSA ≤ 150; HBD ≤ 5; (clogD or AlogP) ≤ 5
Compounds with (clogD or AlogP) > 3 and substructure “Flag”
Source-set size (after exclusions): 832k

Clustering
Pipeline pilot: ECFP6 fingerprints, AvgNumberPerCluster = 20, MaximumDistance = 0.65, MaximumDissimilary for center selection, recenter twice to minimize the maximum distance
46,645 clusters and 6,733 singletons

Class limits
Properties calculated in pipeline pilot
A: 250 ≤ MW ≤ 400; ROT ≤ 5; 50 ≤ PSA ≤ 150; HBD ≤ 3; clogD or AlogP ≤ 3 & no substructure match
B: 250 ≤ MW ≤ 450; ROT ≤ 7; 10 ≤ PSA ≤ 150; HBD ≤ 4; clogD or AlogP ≤ 3 & no substructure match
C: 250 ≤ MW ≤ 500; ROT ≤ 10; PSA ≤ 150; HBD ≤ 5; clogD or AlogP ≤ 3 & no substructure match
D: 250 ≤ MW ≤ 500; ROT ≤ 10; PSA ≤ 150; HBD ≤ 5; clogD ≤ 5 or (A, B, C property limits) with substructure “Flag”
In addition, test fCsp3 if MW > 350
For classes A, B, and C: fCsp3 for MW >350 has to be > 0.2. In class C, it can be <= 0.2 only if ROT >= 2

Selection of “single best” from class-D
Preference of no substructure match over substructure “Flag” and preference of substructure “Flag” over “flat” compounds; “flat” compounds have (fCsp3 <= 0.2 and ROT <= 1). With equivalent substructure status, pick physical chemical property space from class A > B > C

Coverage
Increased coverage for clusters where selection can be done from all-A: 0.5* square root of group size
If sampling needs to happen from class B and/or C: 0.3*square root of group size
Note: group size is count of compounds in cluster after removing all class-D

Diversity assessment of list of confirmed hits

The BCD set in the bacteriology case study yielded 40 confirmed hits. The table below shows a diversity assessment using Bemis-Murcko (BM) scaffolds[18] and level-3 scaffolds (from the scaffold-tree algorithm[26]) for the 40 BCD hits.

Compound set	# Compounds	# Compounds with BM scaffold	# BM scaffolds (total)	# BM scaffolds (shared with partner)	# Compounds with level-3 scaffold	# level-3 scaffolds (total)	# level-3 scaffolds shared with partner
BCD	40	40	37		25	21	

Case study details - Target-1

Exclusions
Substructure “Out”
Frequent hitters: experimentally screened at least 100 times and “hit” at least 1/3 of the time
Anything with properties not in this range: 200 ≤ MW ≤ 700; -1 ≤ clogD ≤ 6
Source-set size (after exclusions): 1,157k

Clustering
ICM MolCluster KNIME node; cluster-size set at 40
29,863 clusters, 6416 singletons

Class limits
A: Complement set (note that docking was done with ICM and pharmacophore matching with MOE)
B: 200 ≤ MW ≤ 400; -1 ≤ clogD ≤ 3; fCsp3 ≥ 0.25 & no substructure match
C: 200 ≤ MW ≤ 500; -1 ≤ clogD ≤ 5 & no substructure match
D: 200 ≤ MW ≤ 700; -1 ≤ clogD ≤ 6 or (B, C property limits) with substructure “Flag”

Selection of “single best” from class-D
Preference of no substructure match over substructure “Flag”; with equivalent substructure status, pick lowest clogD

Coverage
0.7 * square root of group size
Note: group size is count of compounds in cluster after removing all class-D

Diversity assessment of list of confirmed hits

The screening campaign in the target-1 case study yielded 412 confirmed hits. The table below shows a diversity assessment using BM scaffolds and level-3 scaffolds (from the scaffold-tree algorithm) for the 119 BCD hits compared to the 293 hits from the Docking set.

Compound set	# Compounds	# Compounds with BM scaffold	# BM scaffolds (total)	# BM scaffolds (shared with partner)	# Compounds with level-3 scaffold	# level-3 scaffolds (total)	# level-3 scaffolds shared with partner
Docking	293	293	217	4 (BCD)	159	137	2 (BCD)
BCD	119	119	108	4 (Docking)	50	50	2 (Docking)

Case study details - Target-2

Exclusions
Substructure “Out” and “Flag”
Anything with properties not in this range: 250 ≤ MW ≤ 500; ROT ≤ 10; 20 < PSA ≤ 150; HBD ≤ 5; (clogD or AlogP) ≤ 4.5
Compounds without a Nitrogen atom
Source-set size (after exclusions): 736k

Clustering
ICM MolCluster KNIME node, average cluster size 35: 23,673 clusters & 5,461 singletons

Class limits
A: Complement set
B: 250 ≤ MW ≤ 400; ROT ≤ 5; 50 ≤ PSA ≤ 150; HBD ≤ 3; (clogD or AlogP) ≤ 3; fCsp3 ≥ 0.2
C: 250 ≤ MW ≤ 450; ROT ≤ 7; 20 ≤ PSA ≤ 150; HBD ≤ 4; (clogD or AlogP) ≤ 3
D: 250 ≤ MW ≤ 500; ROT ≤ 10; 20 < PSA ≤ 150; HBD ≤ 5; (clogD or AlogP) ≤ 4.5

Selection of “single best” from class-D
Random

Coverage
Increased coverage for clusters where selection can be done from all-B: 0.4 * square root of group size
If sampling needs to happen from class C: 0.2*square root of group size
Note: group size is count of compounds in cluster after removing all class-D

Diversity assessment of list of confirmed hits

The screening campaign in the target-2 case study yielded 232 confirmed hits. The table below shows a diversity assessment using BM scaffolds and level-3 scaffolds (from the scaffold-tree algorithm) for the 81 BCD hits compared to the 151 hits from the three subsets that made-up the core-set.

Compound set	# Compounds	# Compounds with BM scaffold	# BM scaffolds (total)	# BM scaffolds (shared with partner)	# Compounds with level-3 scaffold	# level-3 scaffolds (total)	# level-3 scaffolds shared with partner
Pre-plated diversity	73	72	72	0	61	59	0
Prior hits	10	10	10	0	10	10	1 (BCD)
Privileged scaffolds	68	68	64	1 (BCD)	66	56	1 (BCD)
BCD	81	81	81	1 (PrivScaf)	76	76	1 (PriorHits), 1 (PrivScaf)

Case study details - Malaria

Exclusions
None
Source-set size: 306k

Clustering
ICM MolCluster KNIME node, average cluster size 10 => 21,871 clusters & 8,685 singletons

Class limits
A: Complement set
B: 250 ≤ MW ≤ 400; -1 ≤ SlogP ≤ 3; ROT≤ 5 & no substructure match
C: 200 ≤ MW ≤ 600; -1 ≤ SlogP ≤ 5; ROT≤ 10 & substructure "Flag" (1 match in RDKit PAINS set)
D: MW > 0

Selection of "single best" from class-D
Lowest SlogP (calculated lipophilicity)

Coverage
0.5 * square root of group size
Note: group size is count of compounds in cluster after removing all class-D

Diversity assessment of list of confirmed hits

The BCD screening campaign in the Malaria case study would have yielded 36 confirmed hits, using the annotations from the literature reference[23] for 172 confirmed and cross-validated hits. The table below shows a diversity assessment using BM scaffolds and level-3 scaffolds (from the scaffold-tree algorithm) for the 31 BCD hits compared to the 5 hits from the Similarity set, and the remaining 136 hits in the list of 172 that were not part of this first screening set. Percentage-wise, the BCD hits are 18.0% of the whole confirmed hit-list list, and they contribute 34.1% and 34.9% of all the BM scaffolds and level-3 scaffolds respectively.

Compound set	# Compounds	# Compounds with BM scaffold	# BM scaffolds (total)	# BM scaffolds (shared with partner)	# Compounds with level-3 scaffold	# level-3 scaffolds (total)	# level-3 scaffolds shared with partner
Similarity	5	5	2		5	2	
BCD	31	31	30	11 (Other)	24	22	11 (Other)
Other from confirmed hits	136	136	67	11 (BCD)	125	50	11 (BCD)
All	172	172	88		154	63	

Comparison between BCD, unbiased diversity, and random selections

It is important to emphasize that the BCD design is aimed at increasing the number of quality hits and that total hit-rate is not a driver. A comparison of a BCD design simulation with simulations of an unbiased diversity design and of a random selection supports the statement that BCD is more effective in selecting quality hits. For this comparison, we analyzed an HTS dataset measuring bacterial growth inhibition of an efflux-deficient *E. coli* strain, which had a 790k compound overlap with the source set from the bacteriology case study presented in Table-1.

Table S1-A shows numbers for all compounds in the screening set (single point assay run at 50 µM compound concentration) that overlapped with the source set from the bacteriology case study. The number of primary hits and hit-rate are determined using a cutoff of ≥ 70% growth inhibition. Compounds are categorized according to the classes in the bacteriology case study. In this dataset, the primary hit-rate for the class of high quality compounds (class-A) is 1.4%, which is lower than the overall primary hit-rate (5.4%), and lower than the primary hit-rate in the least desirable class of compounds (9.9% for class-D: compounds with undesirable substructures and/or clogD >3).

We ran the BCD workflow on the 790k source set, using the same settings as were used for the bacteriology case study (table S1-B). We also ran an unbiased diversity selection as a comparison, using the same clustering and coverage criteria, and using the RDKit Diversity Picker without considering class membership (table S1-C). The BCD selection contained almost 3 times more primary hits from class-A compared to the unbiased diversity selection (859 *vs* 311). The class-based primary hit-rates are all comparable to the class-based primary hit-rates in the full-deck dataset, but the overall primary hit-rate for BCD is about half that of the unbiased diversity protocol (2.8% vs 5.8%) because the BCD design picked fewer compounds from the higher hit-rate classes. The same conclusions are true for a comparison of the BCD selection with a random selection, which was created using the KNIME “Random Numbers Generator” node (developed by Vernalis) to pick 71k random numbers out the 790k source set (table S1-D).

In summary, BCD would have resulted in a larger number of high quality primary hits but a lower total number of primary hits compared to an unbiased diversity selection or a random selection if applied to this screening campaign. In cases where the class of high quality compounds has a lower hit-rate, applying BCD means accepting that increasing the number of high quality hits comes at the expense of decreasing the total number of hits.

Table S1. Comparison of BCD, unbiased diversity selection, and random selection using a bacterial growth inhibition dataset

Table S1-A: Full deck HTS

Class	# Compounds in Screening set	Class % in Screening set	# Primary hits	Class % in hit-list	Hit-rate
A	205,822	26	2,820	7	1.4
B	158,285	20	3,258	8	2.1
C	81,032	10	2,570	6	3.2
D	344,481	44	34,047	80	9.9
Total	789,620		42,695		5.4

Table S1-B: BCD selection

Class	# Compounds in Screening set	Class % in Screening set	# Primary hits	Class % in hit-list	Hit-rate
A	50,206	71	859	43	1.7
B	10,216	14	254	13	2.5
C	1,611	2	73	4	4.5
D	8,871	13	813	41	9.2
Total	70,904		1,999		2.8

Table S1-C: Unbiased diversity selection

Class	# Compounds in Screening set	Class % in Screening set	# Primary hits	Class % in hit-list	Hit-rate
A	19,256	27	311	7	1.6
B	13,849	19	284	7	2.1
C	5,819	8	199	5	3.4
D	32,955	46	3,395	81	10.3
Total	71,879		4,189		5.8

Table S1-D: Random selection

Class	# Compounds in Screening set	Class % in Screening set	# Primary hits	Class % in hit-list	Hit-rate
A	18,410	26	250	6	1.4
B	14,094	20	297	8	2.1
C	7,470	11	238	6	3.2
D	31,026	44	3,094	80	10.0
Total	71,000		3,879		5.5

Chemical space analysis

Coverage of chemical space was also assessed using Optibrium's StarDrop Chemical Space tool for the three case studies that had core-sets to complement.[27] The confirmed hits for all three projects were combined, including the hits from the Malaria case study that had not been selected in the design. A chemical space was created for the resulting set of 816 compounds using the "visual clustering" option in Stardrop. A plot trellised by case study is shown in Figure S2 below.

The knowledge-driven sets with explicit target-focus include the docking set for target-1, the prior hits for target-2, and the similarity set for Malaria. For both target-2 and Malaria, the knowledge sets are narrowly targeted and only sample a limited portion of the chemical space. The BCD sets clearly complement and extend the coverage of chemical space around the blue sets. In target-2, the pre-plated diversity set and the BCD set sample a similar extent of space but they still complement each other. The privileged scaffold set of target-2 also samples a wide extent of space. This set includes a peptide-mimetic subset, which is broadly targeted, and expected to be quite diverse. For Malaria, the compounds that were confirmed hits in the literature, but that were not selected in the design (grey) show two clusters that are not sampled by the BCD design. Further inspection shows that those are class-C and class-D compounds that belong to large clusters for which the design selected representatives that did not confirm as hits.

The target-1 plot shows that the two subsets in the design are complementary, and both sample quite a large extent of chemical space. More detail on the design of the docking set explains why that set has such large coverage: the original docking and pharmacophore workflow was set-up to cast a wide net of chemical matter, and was therefore broadly targeted. A set of just over 100k available compounds were flagged as being able to dock in the active site and provide a match to at least one of 4 sparse pharmacophore models. That set of 100k matches was then prioritized using the BCD workflow to pick representatives with the most attractive physical chemical properties and lacking undesirable substructures. That selected docking set was the core around which the BCD workflow selected a structurally diverse complement set. The complement function prevents the selection of more compounds in the already densely sampled area in the right-hand side of the target-1 space.

Note that the TOC graphic has the target-1 plot where the compounds are colored by class, which means that the green set below is subdivided into class B (green), class C (yellow) and class D (red); the blue set is still all blue (all class A).

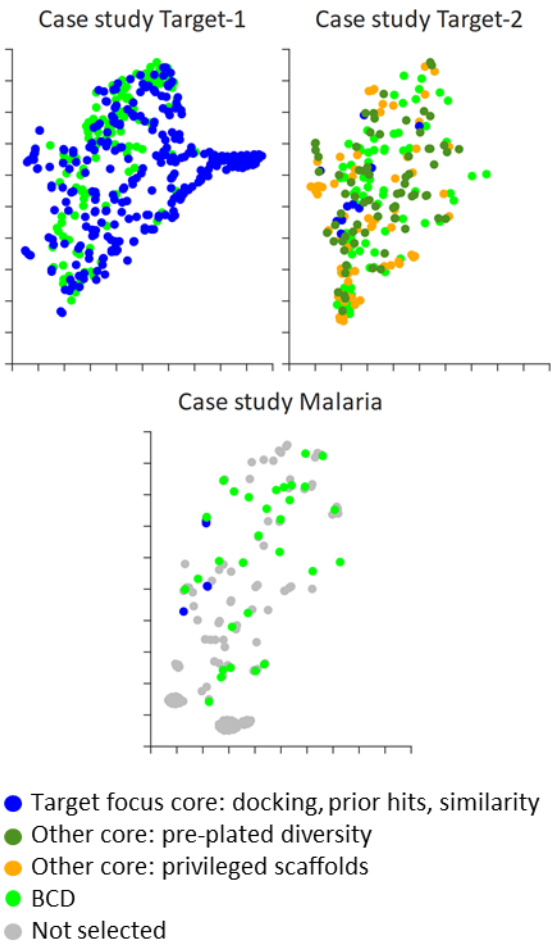

Figure S2. Chemical space coverage of designed screening sets for the three case studies with a core-set