

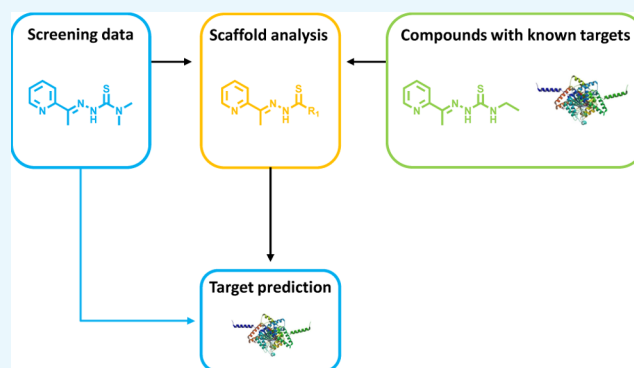
Application of a New Scaffold Concept for Computational Target Deconvolution of Chemical Cancer Cell Line Screens

Ryo Kunimoto, Dilyana Dimova, and Jürgen Bajorath*¹

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Supporting Information

ABSTRACT: Target deconvolution of phenotypic assays is a hot topic in chemical biology and drug discovery. The ultimate goal is the identification of targets for compounds that produce interesting phenotypic readouts. A variety of experimental and computational strategies have been devised to aid this process. A widely applied computational approach infers putative targets of new active molecules on the basis of their chemical similarity to compounds with activity against known targets. Herein, we introduce a molecular scaffold-based variant for similarity-based target deconvolution from chemical cancer cell line screens that were used as a model system for phenotypic assays. A new scaffold type was used for substructure-based similarity assessment, termed analog series-based (ASB) scaffold. Compared with conventional scaffolds and compound-based similarity calculations, target assignment centered on ASB scaffolds resulting from screening hits and bioactive reference compounds restricted the number of target hypotheses in a meaningful way and lead to a significant enrichment of known cancer targets among candidates.



1. INTRODUCTION

Drug discovery research is experiencing a renaissance of phenotypic approaches.^{1,2} Especially high-content and phenotypic screening assays have become a hot topic in recent years.^{3,4} It is generally thought that phenotypic screens might produce leads that are more relevant for addressing complex biology *in vivo* than other compounds identified in target-based assays. Whether or not such expectations might generally be true remains to be determined. Be that as it may, phenotypic discovery is challenged by the need to identify—or at least narrow down—cellular targets for compounds with interesting phenotypic readouts, a process often referred to as target deconvolution.^{5,6} For compound selection and optimization as well as late-stage preclinical evaluation, target knowledge continues to be required in many cases, regardless of how candidate compounds have originally been identified. In addition, there is strong scientific interest in identifying target(s) whose inhibition in cellular environments might result in interesting functional effects.

For target deconvolution from phenotypic screens, different experimental approaches have been developed or adapted,^{5–7} including, among others, various proteomics techniques and the use of small molecular probes with confirmed activity against selected targets. Moreover, target identification has also become an attractive task for computational analysis using different methods. For example, drug-target networks^{8,9} establish compound-based links between targets and help to better

understand complex interactions involving multiple compounds and targets. For drugs, new targets can often be proposed on the basis of network representations that might rationalize side effects.⁹ Such networks can also be generated for bioactive compounds other than drugs and can be computationally analyzed. Furthermore, machine-learning models combining small molecule and target information (e.g., chemical descriptors and protein sequences) have been generated to predict novel compound-target pairings.^{10,11} Moreover, targets of novel active compounds are often inferred from molecular similarity between these compounds and known actives.^{12–14} For similarity calculations, a variety of chemical descriptors and functions exist.^{15,16} Target hypotheses for new chemical entities can be derived not only by molecular similarity calculations producing numerical values but also by assessing substructure relationships between compounds as a measure of similarity. For example, targets can be predicted for new active compounds by identifying structural analogues and comparing their target annotations¹⁷ or on the basis of molecular scaffolds,¹⁸ which are generated to capture core structures of compounds.¹⁹ As such, scaffolds often represent a series of known active compounds sharing the same core. A systematic scaffold analysis provides a structural organization scheme, and

Received: February 24, 2017

Accepted: April 6, 2017

Published: April 14, 2017

target annotations of compounds containing the same scaffold can be assigned to each scaffold.¹⁹ This approach generates activity-annotated scaffold libraries to which new active compounds without known targets can be mapped. If scaffolds of new actives match the existing ones, target hypotheses can be inferred. The classical way of defining scaffolds for medicinal chemistry applications is according to Bemis and Murcko, which gave rise to BM scaffolds.²⁰ These scaffolds are obtained from compounds by the removal of all R-groups while retaining the ring systems and linker fragments connecting the rings.²⁰ Various extensions of the BM scaffold concept such as the Scaffold Tree²¹ have been introduced. The Scaffold Tree decomposes BM scaffolds along tree branches according to chemical rules until only individual rings remain and thereby establishes structural relationships between the scaffolds.²¹ Herein, we report the application of a new scaffold concept termed analog series-based (ASB) scaffold²² to computationally assign potential targets to hits from cancer cell line screens, which are a major resource for phenotypic discovery.²³

2. RESULTS AND DISCUSSION

2.1. ASB Scaffold Concept and Substructure-Based Similarity Assessment. Figure 1 compares the generation of

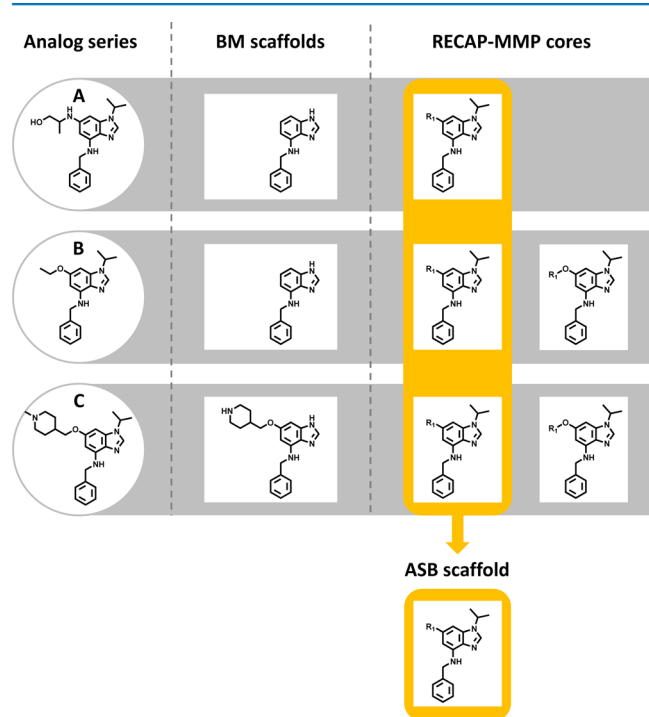


Figure 1. Generation of ASB and BM scaffolds. For a compound series (A–C), the generation of ASB and BM scaffolds is illustrated. Two unique BM scaffolds were isolated from these compounds by removing substituents. RECAP-MMP cores of compounds A–C are shown. The core shared among all compounds (highlighted in orange) represents the ASB scaffold.

ASB and BM scaffolds. Compared with conventional scaffolds, ASB scaffolds were designed to further increase the medicinal chemistry relevance by (i) omitting a formal hierarchical distinction of ring systems, linkers, and substituent; (ii) representing a series of analogues (rather than individual compounds); and (iii) incorporating reaction rules.²² The definition of ASB scaffolds is thus more inclusive and restrictive

than compound-based scaffold concepts. From an ASB scaffold, all analogues of the corresponding series can be regenerated following retrosynthetic rules. The ASB scaffold contains all substructures that are conserved within a series and a consensus substitution site where R-groups distinguish different analogues comprising the series.

For a substructure-based similarity assessment, all compounds represented by the same (BM or ASB) scaffold were assigned to the scaffold and classified as similar. For a compound-based similarity evaluation, pairwise Tanimoto coefficient values for the chosen reference (ChEMBL) and query compounds (from NCI screens) were calculated and a similarity threshold was applied.

2.2. Analysis Concept and Protocol. A major goal of our analysis was the evaluation of a new scaffold concept for the assignment of potential targets to hits from cancer cell line screens. This setup served as a model system for target deconvolution from phenotypic assays. The underlying idea was that structurally very similar active compounds are likely to share targets (which is well-appreciated in medicinal chemistry). Therefore, analog series were systematically extracted from combined screening and ChEMBL compounds to comprehensively capture structural relationships, and the resultant ASB scaffolds were collected. ASB scaffolds representing both screening hits and ChEMBL compounds were prioritized, and known target annotations of ChEMBL compounds were assigned to hits sharing the same ASB scaffold. Then, target annotations were collected for each cell line. The analysis was centered on ASB scaffolds to ensure that only close structural analogues were considered for target transfer from known bioactive compounds to hits. As such, ASB scaffolds provided a “meta structure” for target deconvolution. The analysis protocol that was systematically applied to all 73 cell line screens is illustrated in Figure 2.

The approach is conceptually based on molecular similarity to derive compound-target hypotheses, specifically on substructure-based similarity; that is, compounds are classified as similar if they are represented by the same scaffold. Accordingly, we have compared ASB scaffolds and conventional BM scaffolds in the same analysis context and, in addition, carried out conventional similarity searching as another reference calculation. In the latter case, screening hits were used as templates for similarity searching in ChEMBL. If similar compounds were identified, their target annotations were assigned to the hits.

For our analysis, many properties assigned to scaffolds such as promiscuity, selectivity, or privileged substructure characteristics that are often discussed in medicinal chemistry¹⁹ are not relevant. Neither do we need to consider relative contributions of core structures and R-groups to biological activity. Rather, in the context of our analysis, the use of scaffolds for the structural organization of active compounds becomes critically important, which is only one of many aspects often considered in the scaffold-based analysis of compound activity data.¹⁹

2.3. Scaffold and Compound Statistics. Our analysis protocol identified 99 unique ASB scaffolds shared by screening hits and ChEMBL compounds, 927 shared BM scaffolds, and 25 390 ChEMBL compounds classified on the basis of similarity searching as being similar to screening hits (Table 1). Hence, there were many more compound-based BM than ASB scaffolds and many more similar compounds than scaffolds. For shared ASB and BM scaffolds, 7–40 and 56–388 scaffolds were obtained per cell line screen, with a mean of

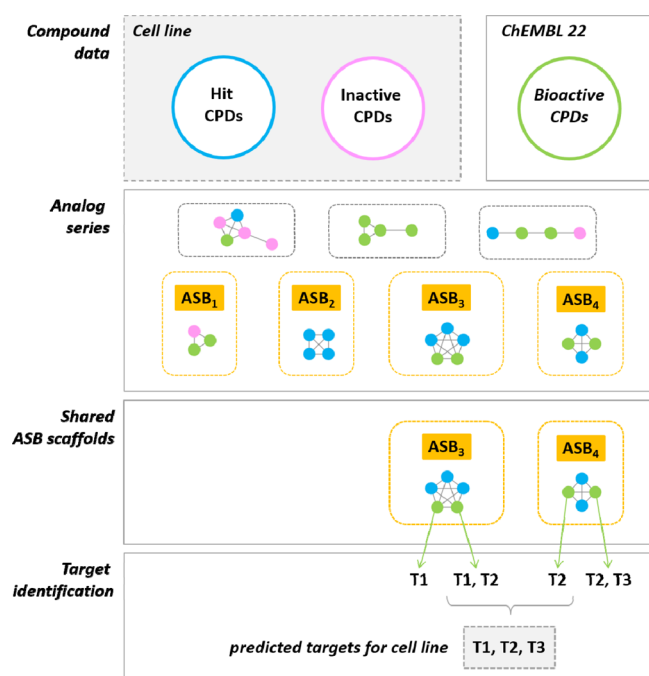


Figure 2. Analysis scheme. For a given cell line, screening compounds (hits, colored in blue; inactive compounds, pink) and bioactive compounds from ChEMBL (green) were pooled. From this compound pool, analog series were extracted (depicted as clusters) and series yielding ASB scaffolds (orange) identified. ASB scaffolds resulting from series containing screening hits and ChEMBL compounds (i.e., ASB₃ and ASB₄) were determined. Target annotations of all bioactive compounds represented by the shared ASB scaffolds were assembled and the union of these targets (i.e., T1, T2, and T3) was assigned to screening hits of this cell line.

Table 1. Scaffold and Similarity Search Statistics^a

	per cell line		TOTAL
	MIN–MAX	AVG	
ASB Scaffolds			
# shared ASB scaffolds	7–40	18.8	99
# targets	30–119	73.7	232
# cancer targets	14–62	26.5	108
cancer target rate (%)	23.3–59.8	36.4	46.6
BM Scaffolds			
# shared BM scaffolds	56–388	209.7	927
# targets	595–1030	925.1	1130
# cancer targets	197–303	275.9	330
cancer target rate (%)	29.0–34.0	30.0	29.2
Similarity Search			
# similar ChEMBL CPDs	962–9465	4883	25 390
# targets	393–972	756.8	1249
# cancer targets	147–311	264.1	366
cancer target rate (%)	31.1–39.5	34.8	34.1

^aThe table reports statistics for scaffold analysis and similarity searching. For ASB and BM scaffolds, ranges (MIN–MAX), averages (AVG), and total numbers (TOTAL) of scaffolds from screening hits and scaffolds that were shared with ChEMBL reference compounds, corresponding targets, and cancer targets are provided across all 73 cell lines. For similarity search calculations, ranges, averages, and total numbers are reported for similar compounds, all targets, and cancer targets.

18.8 and 209.7, respectively (Table 1). Thus, many scaffolds were detected multiple times in different cell line screens. In

addition, the number of similar compounds per cell line ranged from 962 to 9465, with a mean of 4883.

Exemplary shared ASB scaffolds are shown in Figure 3 together with the compound series from which they originated. These examples illustrate another important aspect of the ASB scaffold analysis. In these cases, close screening compound analogues were detected that were either active or inactive in the cell line screen, thus providing immediate opportunities for reassessing assay results by retesting selected hits and/or inactive compounds, prior to the target analysis. In many other instances, shared ASB scaffolds represented only active compounds, as illustrated in Figure 2.

2.4. Target Assignment. **2.4.1. Global Target Distribution.** For each cell line screen, the union of targets associated with shared scaffolds was determined. The 927 shared BM scaffolds yielded a total of 1130 unique targets across all cell lines, with a range of 595 to more than 1000 targets per line, as reported in Table 1. Thus, on the basis of BM scaffolds, approximately 70% of all investigated human targets were assigned to screening hits as potential targets. Similarity searching suggested a larger number of unique 1249 targets of screening hits. However, when ranges of targets over cell line screens were considered—instead of total numbers of unique targets—BM scaffold analysis yielded more targets than similarity searching, with an average 925 versus 756 targets per cell line, respectively (Table 1). Thus, on the basis of compound similarity, individual targets were much less frequently detected than on the basis of shared BM scaffolds. For similarity searching, the number of similar compounds and the resultant targets might be reduced by further increasing the similarity threshold value. Regardless, the control calculations showed that generally applied compound similarity criteria would not be suitable for target assignment across cell line screens. At face value, implicating approximately 70% or more of all preselected targets in activity signals from cell line screens—on the basis of BM scaffolds or compound similarity—was considered not realistic, despite variations observed across different cell lines. By contrast, the structurally more conservative ASB scaffold approach involving multiple compounds significantly reduced the number of target assignments. On the basis of 99 identified shared ASB scaffolds (approximately an order of magnitude less than shared BM scaffolds), a total of 232 unique targets were assigned, with a mean of 74 targets per cell line. Thus, shared ASB scaffolds implicated only approximately 14% of all targets in cell line screens and also controlled the number of targets per line.

2.4.2. Cancer Targets. To specifically focus observed differences in target distributions on the cancer cell line screening, the assignment of known cancer targets was analyzed, which represented a subset of all monitored targets. ASB scaffolds, BM scaffolds, and similarity searching identified 108, 330, and 366 known cancer targets, respectively, as potential targets for screening hits across all cell lines, with ranges of 14–62 (ASB), 197–303 (BM), and 147–311 (similarity) cancer targets per line (Table 1). With one exception (macrophage colony stimulating factor receptor; CSF1R; ChEMBL TID 1844), the set of targets identified on the basis of ASB scaffolds overlapped with the other sets. Table S2 reports the cancer targets assigned on the basis of ASB scaffolds to each cell line screen.

ASB scaffolds assigned approximately one-third of cancer targets compared with BM scaffolds and similarity searching, although the number of all targets differed by more than one

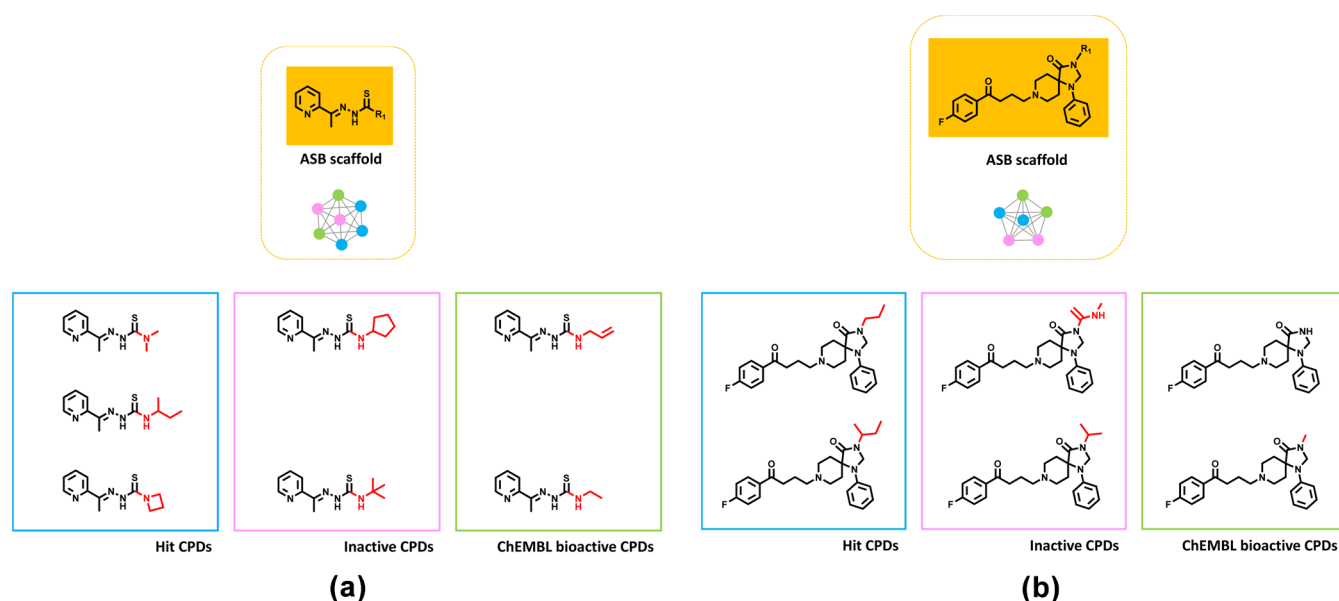


Figure 3. Shared ASB scaffolds. Examples of shared ASB scaffolds (orange background) are shown for (a) SNB-75 (CNS cancer) and (b) HT-29 (colon cancer) cell lines together with corresponding hits (blue box), inactive compounds (pink), and ChEMBL compounds (green). R-groups distinguishing these analogs are shown in red.

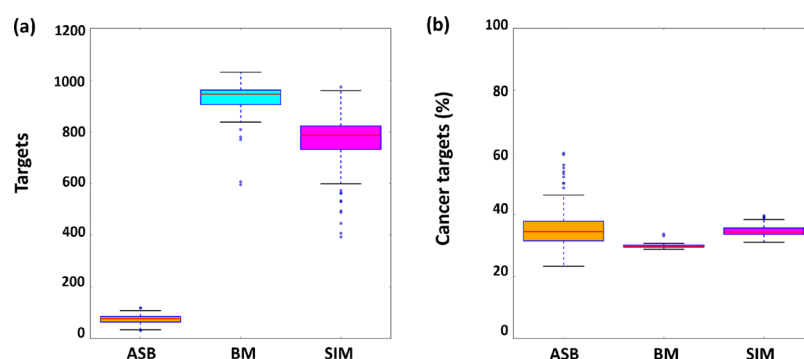


Figure 4. Target distribution. For ASB scaffolds (orange), BM scaffolds (cyan), and similarity searching (SIM, magenta), boxplots report the distribution of (a) all targets and (b) the percentage of cancer targets for all 73 cell lines. Boxplots show the smallest value (bottom), first quartile (lower boundary of the box), median value (red line), third quartile (upper boundary of the box), largest value (top), and outliers (blue dots).

order of magnitude. This corresponded to a significant enrichment of cancer targets among all assigned targets, as illustrated in Figure 4. Although the application of ASB scaffolds resulted in comparably low numbers of assigned targets (Figure 4a), the ratio of cancer targets relative to all targets was higher for ASB than for BM scaffolds and similarity searching (Figure 4b). Given that absolute target numbers were more realistic for ASB than BM scaffolds and similarity searching, the observed enrichment of cancer targets for ASB scaffolds was considered a significant finding. The corroborating evidence for cancer target assignment was provided by the frequent occurrence of established cancer targets across different cell lines, which was clearly evident for ASB scaffolds, given the reduced “target background”. For example, on the basis of ASB scaffolds, well-known cancer targets such as P-glycoprotein 1 and tyrosine-protein kinases Fyn and Src were implicated in 73, 62, and 66 cell line screens, respectively. In total, for ASB scaffolds, 46.6% of all assigned targets were cancer targets, with an average of 36.4% per cell line.

3. CONCLUSIONS

In this work, we have investigated a substructure-based similarity approach to computationally deconvolute targets from 73 chemical cancer cell screens used as a model system for phenotypic assays. Assigning targets on the basis of ligand similarity is a major approach to target identification in phenotypic discovery. The analysis was focused on a recently introduced molecular scaffold definition, ASB scaffolds, designed to further increase the medicinal chemistry relevance of scaffolds as core structure representations. Calculations on the basis of conventional BM scaffolds and whole-molecule Tanimoto similarity served as references. ASB scaffolds are structurally more comprehensive and conservative than other molecular representations for similarity assessment, given their default dependence on compound series. As a consequence, ASB scaffolds produced fewer target hypotheses than BM scaffolds and similarity searching, thereby counteracting the “target inflation” observed for ligand similarity-based target prediction. Moreover, for ASB scaffolds, a significant enrichment of known cancer targets among candidates assigned to screening hits was observed, suggesting that the ASB scaffold

approach provides a promising addition to current computational target deconvolution methods.

4. MATERIALS AND METHODS

4.1. Scaffolds. Conventional BM scaffolds were generated from active compounds by the removal of all R-groups while retaining ring systems and linker fragments connecting rings.²⁰ Furthermore, new ASB scaffolds²² were isolated from compounds. To generate ASB scaffolds, analog series were first systematically identified by applying the matched molecular pair (MMP) approach.²⁴ An MMP is defined as a pair of compounds that are distinguished only by a chemical change at a single site.²⁴ As such, an MMP consists of a conserved MMP core structure and a pair of exchanged substituents. MMPs were generated by applying an algorithm that systematically fragments molecules at exocyclic single bonds and stores resulting cores and substituent fragments in an index table from which MMPs are enumerated.²⁵ Retrosynthetic (RECAP) rules²⁶ were applied to fragment source compounds in which exchanged fragments conform to chemical reactions (thereby replacing random fragmentation steps), yielding RECAP-MMPs.²⁷ From all RECAP-MMPs of active compounds, a network was computed in which nodes represented compounds and edges pairwise RECAP-MMP relationships.²⁸ In this network, each disjoint cluster contained a unique series of analogs²⁸ from which ASB scaffolds were isolated.²² A series of analogs often yielded multiple MMP cores. Therefore, for each series, a computational search was carried out for a core that matched all MMP relationships within the series. If identified, the largest qualifying core then represented the ASB scaffold of the series.²² The generation of ASB scaffolds is computationally efficient as it relies on effective MMP enumeration. Therefore, ASB scaffolds can be generated for large data sets comprising millions of compounds (such as the entire ChEMBL database).²² The generation of BM and ASB scaffolds is schematically illustrated in Figure 1. BM scaffolds were calculated with an in-house implementation using the OpenEye toolkit.²⁹

4.2. Similarity Calculations. As a control for scaffold-based similarity assessment, similarity search calculations were carried out using the extended connectivity fingerprint with bond diameter 4 (ECFP4)³⁰ and a similarity threshold of 0.4 for the Tanimoto coefficient.¹⁶ This threshold value is often used for ECFP4 in virtual compound screening.¹⁶

4.3. Cell Lines and Screening Data. The human tumor cell line growth inhibition assay data from the National Cancer Institute (NCI)³¹ were extracted from PubChem.³² Only compounds screened in confirmatory assays originating from NCI Developmental Therapeutics Program (DTP/NCI) were considered. In total, 2 396 398 assay compounds were screened in 73 cell lines representing 10 different neoplasia (including breast, CNS, colon, leukemia, melanoma, nonsmall cell lung, ovarian, prostate, and renal cancers). Table 2 reports screening statistics for each neoplasia type. Details for all cell lines are provided in Table S1. Assay compounds were designated as active or inactive on the basis of PubChem records. In the following, active compounds are also referred to as hits.

4.4. Reference Compounds. For the scaffold-based similarity analysis, reference compounds were assembled from ChEMBL version 22.³³ Only compounds for which high-confidence activity data were available were considered. Therefore, compounds with direct interactions (type “D”) with human targets at the highest confidence level (ChEMBL

Table 2. Cancer Cell Lines and Screening Data^a

	neoplasia	cell lines	assayed CPDs	active CPDs	inactive CPDs
1	breast	6	161 953	10 031	151 922
2	CNS	8	265 511	13 865	251 646
3	colon	9	310 533	17 070	293 463
4	leukemia	8	231 398	20 082	211 316
5	melanoma	10	360 686	18 693	341 993
6	nonsmall cell lung	11	378 082	19 683	358 399
7	ovarian	7	242 571	12 446	230 125
8	prostate	2	56 284	3195	53 089
9	renal	10	324 513	16 244	308 269
10	small cell lung	2	27 527	1882	25 645

^aThe table provides statistics for the 10 neoplasia types and corresponding screening data. For each neoplasia, the name and number of cell lines are given. In addition, the total number of assayed compounds (CPDs) and the number of active and inactive compounds are reported.

confidence score 9) were selected. Only assay-independent equilibrium constants (K_i values) and assay-dependent IC_{50} values were considered as potency measurements. Approximate measurements (e.g., “>” or “~”) were discarded. If multiple K_i or IC_{50} values were available for the same compound, their geometric mean was calculated as the final potency annotation, provided all values fell within the same order of magnitude. Otherwise, the measurements were discarded. Applying these selection criteria, a total of 224 532 unique compounds were obtained with activity against human 1687 targets.

4.5. Targets. The set of 1687 ChEMBL targets (in the following referred to as targets) was used to assign targets to screening compounds. The subset of known cancer targets was determined. Therefore, known cancer targets were collected from the Therapeutic Target Database,³⁴ and targets implicated in malignant neoplasm were identified on the basis of the ICD-10 code.³⁵ The 1687 ChEMBL targets were found to contain 429 cancer targets.

■ ASSOCIATED CONTENT

📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00215.

Screening details for all cancer cell lines and cancer targets assigned on the basis of ASB scaffolds to each cell line screen (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-2699-306 (J.B.).

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

The study was carried out and the manuscript written with contributions from all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank OpenEye Scientific Software, Inc., for providing a free academic license for the OpenEye toolkit.

REFERENCES

- (1) Swinney, D. C.; Anthony, J. How Were New Medicines Discovered? *Nat. Rev. Drug Discovery* **2011**, *10*, 507–519.
- (2) Lee, J. A.; Uhlik, M. T.; Moxham, C. M.; Tomandl, D.; Sall, D. J. Modern Phenotypic Drug Discovery is a Viable, Neoclassic Pharma Strategy. *J. Med. Chem.* **2012**, *55*, 4527–4538.
- (3) Bickle, M. The Beautiful Cell: High-Content Screening in Drug Discovery. *Anal. Bioanal. Chem.* **2010**, *398*, 219–226.
- (4) Zheng, W.; Thorne, N.; McKew, J. C. Phenotypic Screens as a Renewed Approach for Drug Discovery. *Drug Discovery Today* **2013**, *18*, 1067–1073.
- (5) Terstappen, G. C.; Schlüpen, C.; Raggiaschi, R.; Gaviraghi, G. Target Deconvolution Strategies in Drug Discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 891–903.
- (6) Lee, J.; Bogyo, M. Target Deconvolution Techniques in Modern Phenotypic Profiling. *Curr. Opin. Chem. Biol.* **2013**, *17*, 118–126.
- (7) Schirle, M.; Jenkins, J. L. Identifying Compound Efficacy Targets in Phenotypic Drug Discovery. *Drug Discovery Today* **2016**, *21*, 82–88.
- (8) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. Drug–Target Network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
- (9) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263–266.
- (10) Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. Improved Approach for Proteochemometrics Modeling: Application to Organic Compound–Amine G Protein-Coupled Receptor Interactions. *Bioinformatics* **2005**, *21*, 4289–4296.
- (11) Jacob, L.; Vert, J.-P. Protein–Ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics* **2008**, *24*, 2149–2156.
- (12) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
- (13) Jenkins, J. L.; Bender, A.; Davies, J. W. In Silico Target Fishing: Predicting Biological Targets from Chemical Structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413–421.
- (14) Laggner, C.; Kokel, D.; Setola, V.; Tolia, A.; Lin, H.; Irwin, J. J.; Keiser, M. J.; Cheung, C. Y. J.; Minor, D. L., Jr.; Roth, B. L.; Peterson, R. T.; Shoichet, B. K. Chemical Informatics and Target Identification in a Zebrafish Phenotypic Screen. *Nat. Chem. Biol.* **2012**, *8*, 144–146.
- (15) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
- (16) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204.
- (17) Hu, Y.; Lounkine, E.; Bajorath, J. Many Approved Drugs Have Bioactive Analogs with Different Target Annotations. *AAPS J.* **2014**, *16*, 847–859.
- (18) Hu, Y.; Bajorath, J. Systematic Identification of Scaffolds Representing Compounds Active against Individual Targets and Single or Multiple Target Families. *J. Chem. Inf. Model.* **2013**, *53*, 312–326.
- (19) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59*, 4062–4076.
- (20) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. I. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (21) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
- (22) Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Future Sci. OA* **2016**, *2*, FSO149.
- (23) Moffat, J. G.; Rudolph, J.; Bailey, D. Phenotypic Screening in Cancer Drug Discovery—Past, Present and Future. *Nat. Rev. Drug Discovery* **2014**, *13*, 588–602.
- (24) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.
- (25) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
- (26) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- (27) de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Commun.* **2014**, *5*, 64–67.
- (28) Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, *59*, 7667–7676.
- (29) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2012. <http://www.eyesopen.com>.
- (30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (31) NCI Open Database Compounds, Release 3; National Cancer Institute, 2011.
- (32) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.
- (33) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- (34) Zhu, F.; Shi, Z.; Qin, C.; Tao, L.; Liu, X.; Xu, F.; Zhang, L.; Song, Y.; Liu, X.; Zhang, J.; Han, B.; Zhang, P.; Chen, Y. Z. Therapeutic Target Database Update 2012: A Resource for Facilitating Target-Oriented Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1128–D1136.
- (35) World Health Organization. *The ICD-10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines*; Geneva, SUI, 1992.