Privileged Structures Revisited

Petra Schneider and Gisbert Schneider*

Abstract: *Privileged structures inspire compound library design in medicinal chemistry. We performed a comprehensive analysis of 1.4 million bioactive compounds, with the aim of assessing the prevalence of certain molecular frameworks. We used the Shannon entropy formalism to quantify the promiscuity of the most frequently observed atom scaffolds across the annotated target families. This analysis revealed an apparent inverse relationship between hydrogen-bond-acceptor count of a scaffold and its potential promiscuity. The results further suggest that chemically easily accessible scaffolds can serve as templates for the generation of bespoke compound libraries with differing degrees of multiple target engagement, and heterocyclic, sp³-rich frameworks are particularly suited for target-focused library design. The outcome of our study enables us to place some of the many narratives surrounding the concept of privileged structures into a critical context.*

*I*n 1988, Evans et al. observed that "[…]*certain 'privileged structures' are capable of providing useful ligands for more than one receptor and that judicious modification of such structures could be a viable alternative in the search for new receptor agonists and antagonists.*"[1] Generally speaking, a privileged structure may be considered to possess geometries suitable for decoration with side chains, such that the resulting products bind to different target proteins. Herein, we refer to such molecular frameworks as privileged scaffolds, to avoid confusion with other terminologies. In 2002, we introduced the related concept of frequent hitters[2] for compounds that generate readouts in multiple activity assays.[3–5] Some of these compounds are undesired false positives.[6] Others, however, are truly promiscuous ligands that potently, specifically, reversibly, but not selectively, bind to members of different macromolecular target families (Figure 1).
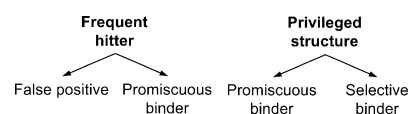

Figure 1. The concepts of frequent hitters and privileged structures.

Privileged scaffolds may be considered desirable for the bespoke design of compound screening libraries.[7] It is important to realize that screening hits from a privileged-scaffold library do not necessarily have to be promiscuous with regard to their target; they can, in fact, represent useful starting points for hit-to-lead expansion, with a well-defined mode of action and selective target engagement. In 2010, Welsch et al. published a broad list of privileged scaffolds compiled from medicinal chemistry literature.[8] Popular examples of molecular frameworks considered promiscuous are indoles, quinolines, coumarines, isoxazolidines, benzimidazoles, thiazolopyrimidines, and arylaminopyrazole.[9] Evidently, the scaffold can contribute to bioactivity directly, via shape or binding interactions, or indirectly, via functionalization potential.[10] Often, the designated privileged scaffolds are inspired by, or derived from, natural products.[11]

In this present study, we have computationally analyzed the scaffold promiscuity of the ChEMBL22[12] compound database. The hypothetical idealized privileged structure, according to Evans' definition, would be found in a ligand that potently interacts with one (selective binder) or many target receptors (promiscuous binder), and with a collection of such compounds containing the privileged structure being able to address all target families. We quantified the ability of a scaffold to interact with members from different target families in terms of its Shannon entropy H [Eq. (1)].[13a] This value corresponds to the divergence between the distribution of the reported activities from the idealized equal distribution across all target families. The Shannon entropy was computed based on the set of compounds containing a certain atom scaffold. It may be considered a set property of the common scaffold.

$$H = -\sum_i p_i \cdot \log_2(p_i), \tag{1}$$

where p_i is the observed fraction of actives for target family i. Information I may be defined as the difference between the maximal and the actual entropy [Eq. (2)].[13b] Accordingly, high values of I designate target-selective scaffolds.

$$I = H_{max} - H. \tag{2}$$

We additionally computed the relative entropy to account for the unequal target activity distribution in ChEMBL22

[*] Dr. P. Schneider, Prof. Dr. G. Schneider
Department of Chemistry and Applied Biosciences
Swiss Federal Institute of Technology (ETH)
Vladimir-Prelog-Weg 4, 8093 Zurich (Switzerland)
E-mail: gisbert@ethz.ch

Dr. P. Schneider
inSili.com LLC
Segantinisteig 3, 8049 Zurich (Switzerland)

⊙ Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/anie.201702816.

[Kullback–Leibler divergence, KLD; Eq. (3)].[13c]

$$KLD = \sum_i p_i \cdot \log_2 \frac{p_i}{q_i}, \qquad (3)$$

where q corresponds to the background (prior) distribution of actives for the different target families considered.

In this present study, ten target families according to the IUPHAR definition[14] with disparate, non-overlapping targets were considered (GPCR, enzyme, kinase, proteinase, nuclear receptor, catalytic receptor, ion channel, transporter, protein, unidentified). The Shannon entropy value for a scaffold is maximal for the equal distribution of activity annotations across the ten target families ($H_{max} = \log_2(10) = 3.32$ bit). Accordingly, the perfectly target-promiscuous compound library would assume this value, while the ideal target-family-selective library would be designated by $H = 0$ bit and $I = 3.32$ bit, respectively.

Database analysis was performed using KNIME[15] workflows with RDkit[16] functions. ChEMBL22 contains 1 397 535 compounds with standardized activity annotations ($K_{d/i/b}$ and EC/IC_{50} values), and a total of 181 888 scaffolds (atom frameworks, "Murcko" scaffolds[17]). By definition, such a scaffold represents the union of all rings and their connecting atoms in a given molecular graph. These atom frameworks do not include side chains. To reduce the risk of

artefacts, we exclusively focused on the most potent compounds (p$Activity \geq 6$, 677 044 compounds) and the most abundant scaffolds with at least 100 compound samples each (585 scaffolds).

The most frequent scaffolds overall were a single phenyl ring (1.6% of all potently active compounds) and acyclic frameworks (0.4%). False-positive spotting according to Rishton[3] and Hann et al.[4] showed most warnings for acyclic scaffolds (up to 37% substructure warnings with $n \geq 2$ flags) and compounds containing a single phenyl ring scaffold (17%). Otherwise, we observed low false-positive potential for these scaffold-focused ligand sets according to the substructure list of Hann et al., with generally more warnings based on the Rishton list (see the Supporting Information).

We subsequently analyzed the activity annotations for all scaffold sets with regard to target-family bias. The quinoline compound set **1a** turned out to possess the most balanced activity spectrum ($H = 2.75$ bit, $I = 0.57$ bit, KLD = 0.37 bit), followed by diphenylmethanes **2a** ($H = 2.54$ bit, $I = 0.78$ bit, KLD = 0.47 bit) and phenylether **3a** derivatives ($H = 2.46$ bit, $I = 0.86$ bit, KLD = 0.87 bit; Table 1). Four of the ten most promiscuous scaffolds (bisphenylether, benzyloxybenzene, phenylbenzamide, benzylindole) are also part of Welsch's list of traditional privileged scaffolds. Overall, among the set of high-entropy scaffolds, we observed a prevalence of small

Table 1: Top-ranking scaffolds according to their promiscuity expressed as Shannon entropy H (**1a–5a**), and according to maximal information content I (**1b–5b**). N = number of potent (p$Activity \geq 6$) compounds containing only the respective atom scaffold and no other ring system.

ID	Scaffold	N	Welsch list[8]	False positives (Rishton)[3]	False positives (Hann et al.)[4]	H (bit)	I (bit)	GPCR	Enzyme	Kinase	Proteinase	Nuclear receptor	Catalytic receptor	Ion channel	Transporter	Protein	Undefined target
1a		285	yes	7%	0%	2.75	0.57	15%	25%	15%	19%	9%	0%	8%	1%	7%	0%
2a		545		23%	4%	2.54	0.78	34%	22%	3%	2%	1%	3%	11%	20%	3%	1%
3a		1184	yes	5%	3%	2.46	0.86	21%	19%	0%	17%	0%	1%	18%	21%	2%	0%
4a		245		1%	18%	2.39	0.93	10%	23%	0%	1%	39%	6%	1%	16%	4%	1%
5a		828	yes	17%	6%	2.38	0.94	23%	35%	0%	11%	0%	0%	19%	9%	2%	1%
1b		280	yes	0%	4%	0	3.32	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2b		282	yes	3%	0%	0	3.32	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%
3b		103		0%	3%	0	3.32	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%
4b		190		0%	0%	0	3.32	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%
5b		870		2%	0%	0	3.32	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%

scaffolds containing two arene rings connected by a short linker (e.g., methyl, ether, amine, amide). Whilst this finding might appear somewhat trivial at first glance, it adds a novel perspective on the concept of privileged structures, coming, as it does, from an entirely target-driven vantage point. Quinn and co-workers recently reported a similar observation for natural products.[18] Furthermore, 12% of the 1822 molecular structures with a molecular weight less than or equal to 2000 g mol^{-1}, which were approved by the U.S. Food and Drug Administration between 1939 and 2016, contain this substructure motif (data retrieved from the e-Drug3D database[19]). Accordingly, it seems prudent to compile general screening compound decks based on scaffolds with low structural intricacy, to increase hit rates and facilitate swift hit-to-lead expansion.[20]

In stark contrast to the architecture of these promiscuous scaffolds, the most information-rich scaffolds retrieved from ChEMBL22 were sp^3-rich heterocyclic molecular frameworks (**1b–5b**, Table 1). In fact, several of them are annotated as target-family selective ($H = 0$ bit, $I = 3.32$ bit, KLD = 1.44–5.14 bit; Table 1, see the Supporting Information). The presence of several of these scaffolds on the Welsh list reveals a dual perception of privileged structures among medicinal chemists. However, according to Evans' original definition of the term,[1] target-selective scaffolds are not "privileged". For some of these examples, the ligand–protein complexes have been solved by X-ray crystallography and exposed directed interactions of the scaffolds with the respective binding pocket, e.g., with kinase hinge residues in the case of scaffold **3b** (PDB ID: 4N70[21]). Hypothesizing different degrees of target promiscuity, it will be worthwhile to comprehensively analyze ligand–protein interaction patterns with regard to the scaffold promiscuity measures introduced in this present study.[22]

A second computational analysis focused on the predicted average promiscuity score of the scaffold-centered ligand sets. In contrast to the scaffold Shannon entropy, this index is computed for each compound individually. We employed a neural network model for this purpose, which we had previously trained to distinguish between undesired (potential false positives) and target-promiscuous frequent hitters.[23] We found a lack of correlation (Pearson $r = 0.05$, 0.03) between the two entropy measures (H, KLD) and the average neural network score. This result shows that high Shannon entropy (low relative entropy) of the scaffold does not necessarily reflect the binding promiscuity of the individual ligands.

Importantly, correlation analysis suggested an inverse relationship between scaffold target promiscuity (H) and both the sp^3 atom count ($r = -0.24$) and the number of hydrogen-bond acceptors ($r = -0.28$). Accordingly, scaffolds with few hydrogen-bond acceptors have a greater promiscuity potential than scaffolds with many such interaction points. This hypothesis was strengthened when focusing on scaffolds with at least 300 active compound samples each ($r_{sp3} = -0.38$; $r_{\text{H-bond acceptors}} = -0.53$; Figure 2). This finding is in line with former studies of the influence of scaffold structure and complexity on hit-to-drug progression.[19, 24] Notably, we did
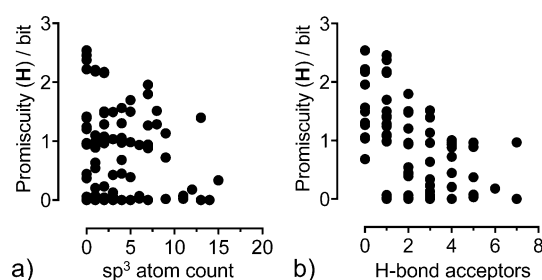

Figure 2. Relationship between the target promiscuity of scaffolds (expressed as Shannon entropy *H*, based on ChEMBL22 activity annotations) and the number of sp^3-hybridised centers (a) and number of hydrogen-bond acceptors (b). Only scaffolds with more than 300 potent ligands were considered (87 scaffolds).

not observe a meaningful correlation between H or KLD and the log$P$ of the compounds ($r = 0.07$).

This present study showcases a straightforward information-theoretical approach for the quantification of scaffold promiscuity, which could assist medicinal chemists in scaffold prioritization and screening library design. However, there are several caveats to keep in mind when interpreting the results presented here. Evidently, the information presented merely reflects the current status of the ChEMBL activity annotations, and there may be many more hitherto-unknown ligand activities.[25] In addition, ChEMBL may not reflect the true diversity of pharmacologically relevant compound structures, since the majority of the proprietary hits and lead compounds from industry are not contained in the ChEMBL database. Although we considered sp^3 hybridization, our analysis does not explicitly account for the three-dimensionality of scaffolds and compounds. It will now be worthwhile to systematically apply the quantitative Shannon entropy concept to analyze the relationship between the shape and flexibility of a scaffold and its promiscuity.[26]

The Shannon concept has been used before to assess chemical scaffold diversity.[27] Our present analysis complements these chemical-diversity analyses by contributing a quantitative target-oriented vantage point. Promiscuity seems to be a property of the full compound, and for most of the published examples (ChEMBL) cannot be attributed to the scaffold alone. More specifically, the results of this present study do not corroborate the existence of an apparent generalizable relationship between the size of a molecular scaffold and the ability of a compound to bind to members from different target families. However, the observed inverse correlation between the number of sp^3-hybridised centers and hydrogen-bond acceptors present in a scaffold and the promiscuity of the respective scaffold-based compound libraries qualifies the use of certain scaffolds and fragments for target-focused hit discovery. Our study further revealed that chemically easily accessible scaffolds can serve as templates for the generation of compounds that could bind to almost all target families (i.e., scaffolds with high Shannon entropy). Together with the neural network score, this information could be utilized to generate custom-made combinatorial screening decks, depending on the intended target(s) or disease.

[1] B. E. Evans, K. E. Rittle, M. G. Bock, R. M. DiPardo, R. M. Freidinger, W. L. Whitter, G. F. Lundell, D. F. Veber, P. S. Anderson, R. S. L. Chang, V. J. Lotti, D. J. Cerino, T. B. Chen, P. J. Kling, K. A. Kunkel, J. P. Springer, J. Hirshfieldt, *J. Med. Chem.* **1988**, *31*, 2235–2246.

[2] a) O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E. M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, *J. Med. Chem.* **2002**, *45*, 137–142; b) G. Schneider, P. Schneider, in: *Chemogenomics in Drug Discovery* (Eds.: H. Kubinyi, G. Müller), Wiley-VCH, Weinheim, **2004**, pp. 341–376.

[3] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382–384.

[4] M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller, N. Ramsden, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.

[5] W. P. Walters, M. Namchuk, *Nat. Rev. Drug Discovery* **2003**, *2*, 259–266.

[6] a) J. Seidler, S. L. McGovern, T. N. Doman, B. K. Shoichet, *J. Med. Chem.* **2003**, *46*, 4477–4486; b) J. Baell, M. A. Walters, *Nature* **2014**, *513*, 481–483; c) J. W. Nissink, S. Blackburn, *Future Med. Chem.* **2014**, *6*, 1113–1126; d) J. L. Dahlin, M. A. Walters, *Future Med. Chem.* **2014**, *6*, 1265–1290; e) J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian, B. K. Shoichet, *J. Med. Chem.* **2015**, *58*, 7076–7087; f) G. Papadatos, A. Gaulton, A. Hersey, J. P. Overington, *J. Comput. Aided Mol. Des.* **2015**, *29*, 885–889.

[7] a) H. Zhao, J. Dietrich, *Expert Opin. Drug Discovery* **2015**, *10*, 781–790; b) Y. Song, W. Chen, D. Kang, Q. Zhang, P. Zhan, X. Liu, *Comb. Chem. High Throughput Screening* **2014**, *17*, 536–553; c) R. W. DeSimone, K. S. Currie, S. A. Mitchell, J. W. Darrow, D. A. Pippin, *Comb. Chem. High Throughput Screening* **2004**, *7*, 473–494.

[8] M. E. Welsch, S. A. Snyder, B. R. Stockwell, *Curr. Opin. Chem. Biol.* **2010**, *14*, 347–361.

[9] a) T. V. Sravanthi, S. L. Manju, *Eur. J. Pharm. Sci.* **2016**, *91*, 1–10; b) M. Berthet, T. Cheviet, G. Dujardin, I. Parrot, J. Martinez, *Chem. Rev.* **2016**, *116*, 15235–15283; c) M. Marinozzi, G. Marcelli, A. Carotti, *Mini-Rev. Med. Chem.* **2015**, *15*, 272–299; d) G. Kaur, M. Kaur, O. Silakari, *Mini-Rev. Med. Chem.* **2014**, *14*, 747–767; e) S. Lal, T. J. Snape, *Curr. Med. Chem.* **2012**, *19*, 4828–4837.

[10] J. J. Yang, O. Ursu, C. A. Lipinski, L. A. Sklar, T. I. Oprea, C. G. Bologa, *J. Cheminf.* **2016**, *8*, 29.

[11] a) T. Rodrigues, D. Reker, P. Schneider, G. Schneider, *Nat. Chem.* **2016**, *8*, 531–541; b) F. R. de Sá Alves, E. J. Barreiro, C. A. Fraga, *Mini-Rev. Med. Chem.* **2009**, *9*, 782–793; c) J. Polanski, A. Kurczyk, A. Bak, R. Musiol, *Curr. Med. Chem.* **2012**, *19*, 1921–1945.

[12] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.

[13] a) C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379–423; b) N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd ed, MIT Press, Cambridge, MA, **1948**, p. 18; c) S. Kullback, R. A. Leibler, *Ann. Math. Stat.* **1951**, *22*, 79–86.

[14] C. Southan, J. L. Sharman, H. E. Benson, E. Faccenda, A. J. Pawson, S. P. H. Alexander, O. P. Buneman, A. P. Davenport, J. C. McGrath, J. A. Peters, M. Spedding, W. A. Catterall, D. Fabbro, J. A. Davies, *Nucleic Acids Res.* **2016**, *44*, D1054–D1068.

[15] M. R. Berthold, N. Cebron, F. Dill, T. R. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, in: *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Heidelberg, **2007**, pp. 319–326.

[16] RDKit: Cheminformatics and Machine Learning Software, 2013, http://www.rdkit.org.

[17] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[18] M. Pascolutti, M. Campitelli, B. Nguyen, N. Pham, A. D. Gorse, R. J. Quinn, *PLoS One* **2015**, *10*, e0120942.

[19] E. Pihan, L. Colliandre, J. F. Guichou, D. Douguet, *Bioinformatics* **2012**, *28*, 1540–1541.

[20] M. M. Hann, A. R. Leach, G. Harper, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.

[21] M. T. Burger, W. Han, J. Lan, G. Nishiguchi, C. Bellamacina, M. Lindval, G. Atallah, Y. Ding, M. Mathur, C. McBride, E. L. Beans, K. Muller, V. Tamez, Y. Zhang, K. Huh, P. Feucht, T. Zavorotinskaya, Y. Dai, J. Holash, J. Castillo, J. Langowski, Y. Wang, M. Y. Chen, P. D. Garcia, *ACS Med. Chem. Lett.* **2013**, *4*, 1193–1197.

[22] a) J. Bajorath, *Mol. Inf.* **2016**, *35*, 583–587; b) S. Glinca, G. Klebe, *J. Chem. Inf. Model.* **2013**, *53*, 2082–2092; c) M. Weisel, J. M. Kriegl, G. Schneider, *ChemBioChem* **2010**, *11*, 556–563; d) M. Weisel, E. Proschak, J. M. Kriegl, G. Schneider, *Proteomics* **2009**, *9*, 451–459; e) R. Abagyan, I. Kufareva, *Methods Mol. Biol.* **2009**, *575*, 249–279.

[23] P. Schneider, M. Röthlisberger, D. Reker, G. Schneider, *Chem. Commun.* **2016**, *52*, 1135–1138.

[24] a) F. Lovering, J. Bikker, C. Humblet, *J. Med. Chem.* **2009**, *52*, 6752–6756; b) J. Y. Ortholand, A. Ganessan, *Curr. Opin. Chem. Biol.* **2004**, *8*, 271–280.

[25] a) Y. Hu, J. Bajorath, *PLoS One* **2015**, *10*, e0126838; b) Y. Hu, J. Bajorath, *Drug Discovery Today* **2013**, *18*, 644–650.

[26] a) J. Meyers, M. Carter, N. Y. Mok, N. Brown, *Future Med. Chem.* **2016**, *8*, 1753–1767; b) G. Müller, T. Berkenbosch, J. C. Benningshof, D. Stumpfe, J. Bajorath, *Chemistry* **2017**, *23*, 703–710; c) A. D. Morley, A. Pugliese, K. Birchall, J. Bower, P. Brennan, N. Brown, T. Chapman, M. Drysdale, I. H. Gilbert, S. Hoelder, A. Jordan, S. V. Ley, A. Merritt, D. Miller, M. E. Swarbrick, P. G. Wyatt, *Drug Discovery Today* **2013**, *18*, 1221–1227; d) M. Wirth, W. H. Sauer, *Mol. Inf.* **2011**, *30*, 677–688.

[27] a) B. B. Yan, M. Z. Xue, B. Xiong, K. Liu, D. Y. Hu, J. K. Shen, *Acta Pharmacol. Sin.* **2009**, *30*, 251–258; b) J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, T. Scior, *QSAR Comb. Sci.* **2009**, *28*, 1551–1560; c) S. R. Langdon, N. Brown, J. Blagg, *J. Chem. Inf. Model.* **2011**, *51*, 2174–2185; d) Y. Hu, A. M. Wassermann, E. Loukine, J. Bajorath, *J. Med. Chem.* **2010**, *53*, 752–758; e) D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, H. Jhoti, *Nat. Rev. Drug Discovery* **2016**, *15*, 605–619.