# Part One
# Scaffolds: Identification, Representation Diversity, and Navigation

*Scaffold Hopping in Medicinal Chemistry,* First Edition. Edited by Nathan Brown.

# 1
# Identifying and Representing Scaffolds

*Nathan Brown*

## 1.1
## Introduction

Drug discovery and design is an inherently multiobjective optimization process. Many different properties require optimization to develop a drug that satisfies the key objectives of safety and efficacy. Scaffolds and scaffold hopping, the subject of this book, are an attempt to identify appropriate molecular scaffolds to replace those that have already been identified [1,2]. Scaffold hopping has also been referred to as lead hopping, leapfrogging, chemotype switching, and scaffold searching in the literature [3–6]. Scaffold hopping is an approach to modulating important properties that may contravene what makes a successful drug: safety and efficacy. Therefore, due consideration of alternative scaffolds should be considered throughout a drug discovery program, but it is perhaps more easily explored earlier in the process. Scaffold hopping is a subset of bioisosteric replacement that focuses explicitly on identifying and replacing appropriate central cores that function similarly in some properties while optimizing other properties. While bioisosteric replacement is not considered to a significant degree in this book, a sister volume has recently been published [7], many of the approaches discussed in this book are also applicable to bioisosteric replacement.

Some properties that can be modulated by judicious replacement of scaffolds are binding affinity, lipophilicity, polarity, toxicity, and issues around intellectual property rights. Binding affinity can sometimes be improved by introducing a more rigid scaffold. This is due to the conformation being preorganized for favorable interactions. One example of this was shown recently in a stearoyl-CoA desaturase inhibitor [8]. An increase in lipophilicity can lead to an increase in cellular permeability. The replacement of a benzimidazole scaffold with the more lipophilic indole moiety was recently presented as a scaffold replacement in an inhibitor targeting N5SB polymerase for the treatment against the hepatitis C virus [9]. Conversely, replacing a more lipophilic core with the one that is more polar can improve the solubility of a compound. The same two scaffolds as before were used, but this time the objective was to improve solubililty, so the indole was replaced for the benzimidazole [10]. Sometimes, the central core of a lead molecule can have

*Scaffold Hopping in Medicinal Chemistry*, First Edition. Edited by Nathan Brown.

pathological conditions in toxicity that needs to be addressed to decrease the chances of attrition in drug development. One COX-2 inhibitor series consisted of a central scaffold of diarylimidazothiazole, which can be metabolized to thiophene S-oxide leading to toxic effects. However, this scaffold can be replaced with diarylthiazolotriazole to mitigate such concerns [11,12]. Finally, although not a property of the molecules under consideration *per se*, it is often important to move away from an identified scaffold that exhibits favorable properties due to the scaffold having already been patented. The definition of Markush structures will be discussed later in this chapter and more extensively in Chapter 2.

Given the different outcomes that lead to what can be called a scaffold hop, one can surmise that there must be different definitions of what constitutes a scaffold hop and indeed the definition of a scaffold itself. This chapter particularly focuses on identifying and representing scaffolds in drug discovery. Markush structures will be introduced as a representation of scaffolds for inclusion in patents to protect intellectual rights around a particular defined core, which will also be discussed in Chapter 2. Objective and invariant representations of scaffolds are essential for diversity analyses of scaffolds and understanding the scaffold coverage and diversity of our screening libraries. Some of the more popular objective and invariant scaffold identification methods will be introduced later in this chapter. The applications of these approaches will be discussed in more detail later in this book, with particular reference to the coverage of scaffolds in medicinal chemistry space.

## 1.2 History of Scaffold Representations

Probably the first description, which is still in common use today, is the Markush structure introduced by Eugene A. Markush from the Pharma-Chemical Corporation in a patent granted in 1924 [13]. Markush defined a generic structure in prose that allowed for his patent to cover an entire family of pyrazolone dye molecules:

> I have discovered that the diazo compound of unsulphonated amidobenzol (aniline) or its homologues (such as toluidine, xylidine, etc.) in all their isomeric forms such as their ortho, meta and para compounds, or in their mixtures or halogen substitutes, may be coupled with halogen substituted pyrazolones (such as dichlor-sulpho-phenyl-carboxlic-acid pyrazolone) to produce dyes which are exceptionally fast to light, which will dye wool and silk from an acidulated bath.

More specifically, Markush's claims were as follows:

1) The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

2) The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.
3) The process for the manufacture of dyes which comprises coupling dichlor-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

Interestingly, the careful reader will note that claims 1 and 2 in Markush's patent are exactly the same. It is not known why this would have been the case, but it may be speculated that it was a simple clerical error with Markush originally intending to make a small change in the second claim as can be seen in the third claim. Therefore, Markush's patent may not have been as extensive since it is possible one of his claims did not appear in the final patent.

Markush successfully defended his use of generic structure definitions at the US Supreme Court, defining a scaffold together with defined lists of substituents on that scaffold. Extending the chemistry space combinatorially from this simple schema can lead to many compounds being covered by a single patent. However, there remains a burden on the patent holders that although it may not be necessary to synthesize every exemplar from the enumerated set of compounds, each of the compounds must be synthetically feasible to someone skilled in the art. A patent may not be defendable if any of the compounds protected by a Markush claim cannot subsequently be synthesized.

An example of a possible Markush structure for the HSP90 inhibitor, NVP-AUY922 (Figure 1.1a) is given in Figure 1.1b. However, an example of a medicinal chemist may determine as the molecular scaffold is given in Figure 1.1c [14,15].

The Markush claim discussed above is clearly a mechanism for extending the protection of a single patent application to a multitude of related and defined compounds. The earliest reference to what we would now call a molecular scaffold definition that this author could identify was in 1969, in an article published in the *Journal of the American Chemical Society*, which provided the following definition [16]:

> The ring system is highly rigid, and can act as a scaffold for placing functional groups in set geometric relationships to one another for systematic studies of transannular and multiple functional group effects on physical and chemical properties.

Clearly, this is a simple description of what constitutes a molecular scaffold and is readily understandable to a scientist active in medicinal chemistry and a specific example of a structural scaffold. However, its simple definition belies an inherent challenge in the identification of molecular scaffolds. Quite often, a medicinal chemist can identify what they would refer to as a molecular scaffold. This often involves identification of synthetic handles. The challenge here though is to

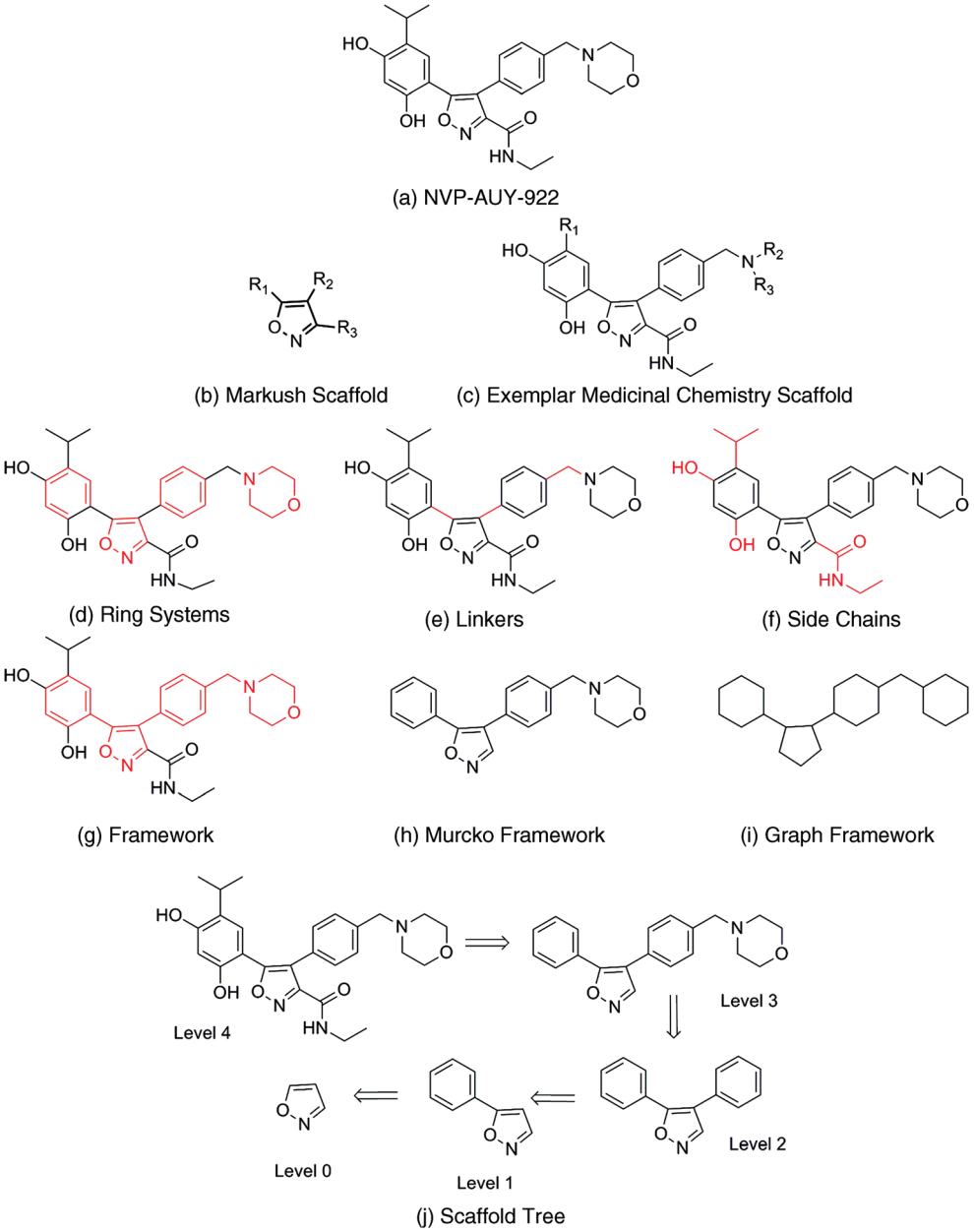


**Figure 1.1** The HSP90 inhibitor NVP-AUY922 depicted using different scaffold representations. (Reproduced from Ref. [20].)

understand how the scaffold has been determined, but this is a soft problem that is not capable of being reduced to an objective and invariant set of rules for scaffold identification. An expert medicinal chemist will bring to bear a wealth of knowledge from their particular research foci during their career and knowledge of synthetic routes: essentially, their intuition. Given a molecule, there are many ways of fragmenting that molecule that may render the key molecular scaffold of interest for the domain of applicability.

## 1.3 Functional versus Structural Molecular Scaffolds

Scaffolds can be divided roughly into two particular classes: functional and structural. A *functional scaffold* can be seen as a scaffold that contains the interacting elements with the target. Once defined, medicinal chemistry design strategies can concentrate on further improving potency while also optimizing selectivity and other properties, such as improving solubility. Conversely, a *structural scaffold* is one that literally provides the scaffolding of exit vectors in the appropriate geometries to allow key interacting moieties to be introduced to decorate the scaffold.

## 1.4 Objective and Invariant Scaffold Representations

It is important to be able define objective and invariant scaffold representations of molecules not only to permit rapid calculation of the scaffold representations but to also allow comparisons between the scaffolds of different molecules. Much research continues into objective and invariant scaffold representations, but here we summarize some of the methods that have seen significant utility. These scaffold representations use definitions of structural components of molecules: ring systems (Figure 1.1d), linkers (Figure 1.1e), side chains (Figure 1.1f), and the framework that is a connected set of ring systems and linkers (Figure 1.1g).

### 1.4.1 Molecular Frameworks

One of the first approaches to generating molecular scaffolds from individual molecules was the *molecular framework* (often referred to as *Murcko frameworks*) and *graph framework* representations [17]. Here, each molecule is treated independently; therefore, the method is objective and invariant.

The molecular framework is generated from an individual molecule by pruning all acyclic substructures that do not connect two cyclic systems (Figure 1.1h). The graph framework is a further abstraction in which the atom labels and bond orders are discarded to provide a simple abstraction of the general topology of the

molecule. The molecular (or Murcko) and graph framework representations of NVP-AUY-922 are given in Figure 1.1h and i, respectively.

This work was the first approach to classifying the crude shapes of molecules in terms of their cyclic frameworks. The inclusion of these topological representations and calculations of equivalences were suggested as being ripe for application to the *de novo* design problem. The study also highlighted the lack of scaffold diversity based on these representations in drug-like molecules and concluded that this would be an area of interest for medicinal chemists to understand which frameworks are underrepresented. The framework definitions were also applied to analyze the scaffold diversity in the Chemical Abstracts Service registry of 24 282 284 compounds at the time of publication in 2008 [18]. This application will be discussed more thoroughly in Chapter 3.

### 1.4.2 Scaffold Tree

Schuffenhauer *et al.* [19] defined the scaffold tree as a set of prioritization rules to systematically prune a given molecule. Starting from the molecular framework defined by Bemis and Murcko [17], rings are sequentially removed using the prioritization rules until only a single ring remains, the so-called level 0 scaffold. The prioritization rules defined for the scaffold tree are provided in Table 1.1.

By application of each of the prioritization rules defined by the scaffold tree method, each molecule in a data set is represented as a directed linear path of iteratively pruned fragments. The scaffold tree pruning strategy is data set independent: a given molecule will always result in the same result. However, the generation of the scaffold tree itself is a summary of a given data set. The pruning path of each molecule in a data set is analyzed and paths merged with one another to generate one or more scaffold trees. For a given data set, one scaffold tree will be the result if all of those molecules in the data set have the same common single

**Table 1.1** The prioritization rules defined to prune ring systems in the generation of the scaffold tree.

| | |
|---|---|
| 1 | Remove three-member heterocycles |
| 2 | Retain macrocycles of greater than 11 members |
| 3 | Remove rings first by longest acyclic linker |
| 4 | Retain spiro, nonlinear, fused and bridged rings |
| 5 | Retain bridged over spiro rings |
| 6 | Remove rings of size 3, 5, and 6 first |
| 7 | Fully aromatic rings should not be removed if remaining system is not aromatic |
| 8 | Remove rings with fewest heteroatoms first |
| 9 | If (8) is equal, use precedence relationship of N > O > S |
| 10 | Remove smaller rings first |
| 11 | Retain saturated rings |
| 12 | Remove rings with a heteroatom connected to a linker |
| 13 | Tiebreaking rule based on alphabetic ordering of a canonical SMILES representation |

ring, the level 0 scaffold. With each additional level of the scaffold tree, the rings are included from each of the molecules in reverse order of the pruning process. Therefore, the level 1 scaffold will typically contain two ring systems (although this is not the case for monocyclic rings).

The advent of the scaffold tree method provided a simple, yet interpretable, hierarchical classification of data sets of molecules using an objective and invariant structural pruning strategy. The authors in their original work postulated a number of applications of the scaffold tree, including the analysis of structure–activity relationships (SAR), particularly in the context of high-throughput screening (HTS) campaigns. The scaffold tree from a pyruvate kinase assay of 602 active and 50 000 inactive molecules is given in Figure 1.2. Analysis of compound collections offered by commercial compound vendors or of the internal compound collection of an organization is an approach to investigating the structural diversity of these libraries, which may or may not be desirous depending on the purpose of those libraries.

In 2011, Langdon *et al.* [20] published a scaffold diversity analysis using the level 1 of the scaffold tree compared with molecular frameworks across the range of compound libraries, including those from vendors, internal fragment and lead-like screening files, exemplified medicinal chemistry from the literature and database of marketed drugs. This work is presented in further detail by the authors of this study in Chapter 3.

The scaffold tree algorithm has more recently been extended to generate Scaffold Networks by some of the original authors of the study [21]. As the name implies, Scaffold Networks generate a highly interconnected network of relationships between molecules and their entire enumerated sets of fragments.

## 1.5 Maximum Common Substructures

The calculation of the maximum common substructure (MCS) of a given congeneric series of molecules is formally not solvable in polynomial time, although it can be approximated in most cases for chemical structures and used effectively [22].

The challenge of using MCS algorithms on congeneric series can be overcome largely by introducing an iterative clustering, based on molecular similarity, followed by application of an MCS algorithm, which iterates until a termination condition is satisfied regarding the quality of the MCS at each stage. Nicolaou *et al.* [23] published the first implementation of an iterative approach to calculating the set of MCSs over a scaffold heterogeneous data set. This iterative approach allowed the generation of MCS groups from large sets of diverse molecules typically found in HTS libraries.

Clark and Labute [24] apply the scaffold tree approach by Schuffenhauer *et al.* for the detection, alignment, and assignment of scaffolds. Once the scaffold tree is generated, a score is generated for each fragment in the tree according to

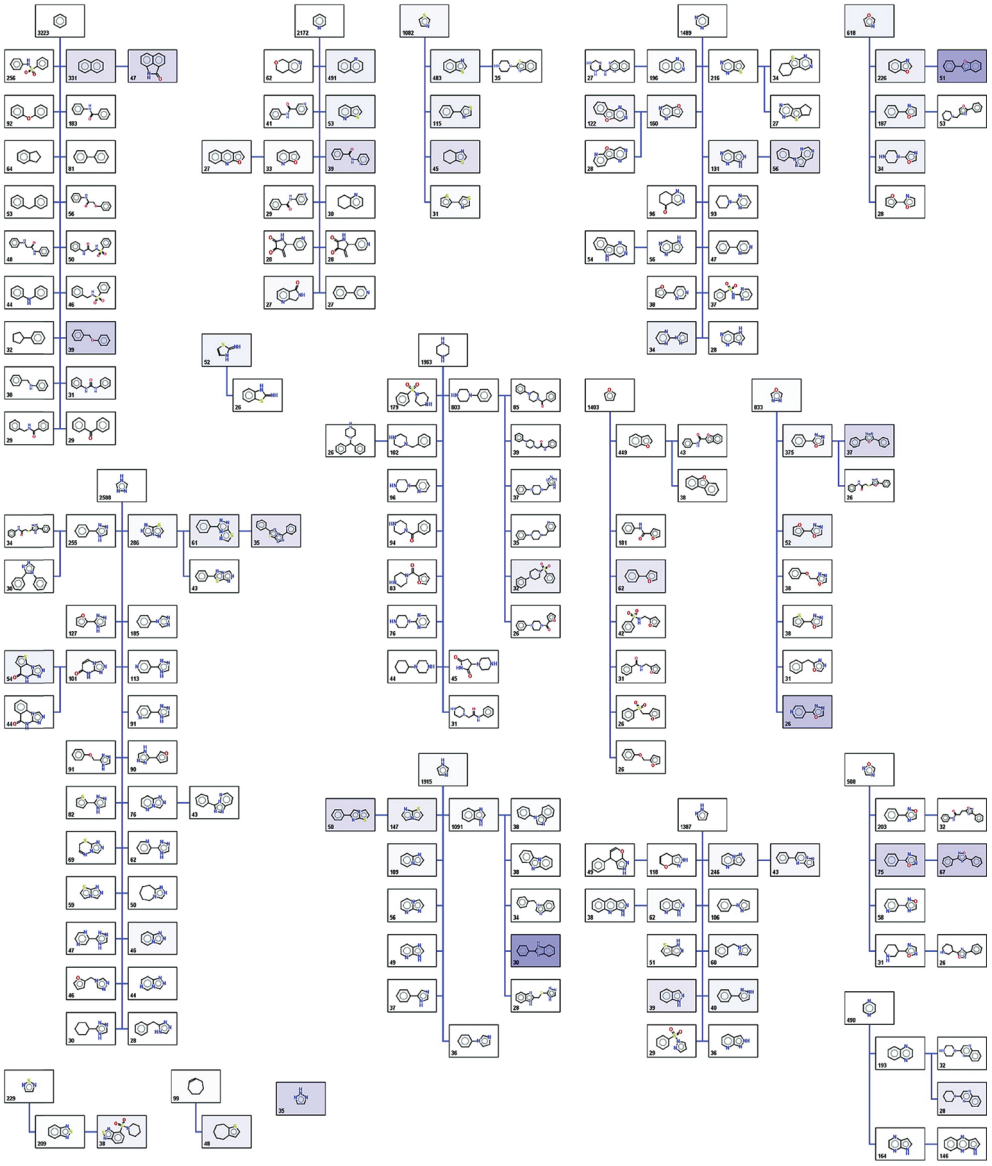

**Figure 1.2** Scaffold tree for the results of pyruvate kinase assay. Color intensity represents the ratio of active and inactive molecules with these scaffolds. (Reproduced from Ref. [19].)

the fraction of the remaining molecule set that contains that fragment, number of heavy atoms in the fragment, theoretical number of R groups, number of fragments selected in previous iterations, and the similarity of the fragment to each previously selected fragment. The method published addresses multiple scenarios of databases with varying degrees of scaffold homogeneity, including

homogeneous single scaffolds, misleading nonscaffolds, multiple similar common scaffolds, ambiguous common scaffolds, symmetrical common scaffolds, overly common scaffolds, and user-specified scaffolds.

While the method developed by Clark and Labute is not an MCS algorithm in principle, the results were demonstrated to be closer to the expectations of a medicinal chemist.

## 1.6 Privileged Scaffolds

A scaffold is deemed to be privileged if it appears many times across multiple targets [25]. Privileged scaffolds were first referred to in 1988 as "privileged structures" [26]. However, the significance of its privilege may not be as a result of commonality in terms of function. Depending on what is decorating an identified scaffold, the function of the resultant molecule with decoration may be significantly different. Take the example of piperazine, which may be monosubstituted or disubstituted, its scaffolding impact can be very different if it is a spiro center or not. It is important to understand the context of the scaffold in terms of biological target and also to realize that a particular scaffold may have been explored more deeply in one medicinal chemistry project than the other for various reasons.

## 1.7 Conclusions

This chapter has introduced a number of, but not exhaustive, published methods for scaffold identification. While it is typically intuitive for an expert medicinal chemist to be able to identify the scaffold of a given molecule, this may not be the same scaffold identified by other similar experts. However, for computational analysis, it is desirable to have an objective and invariant definition of a scaffold. The objective and invariant identification of the molecular scaffold from either an individual molecule or a set of congeneric molecules remains an unsolved problem. This is essentially due to soft issues surrounding scaffold definitions as discussed, but algorithms have been developed that can identify and appropriate scaffold representation in most cases.

This book is structured into three distinct parts. Part One covers different approaches to scaffold representations, analysis of scaffold diversity, and navigating the scaffold space. In this part, concepts discussed briefly here will be expanded upon with more consideration given to Markush structures, analysis of the scaffold diversity, and mining and hopping in these data. Finally, the part concludes with approaches to exploring virtual scaffold spaces that can be enumerated.

Part Two represents a selection of scaffold hopping algorithms and methods that represent a subset of the current state of the art. This part covers methods that

utilize topological representations of molecules, molecular shape, pharmacophores, and explicit information from protein–ligand cocrystal structures.

Part Three includes a selection of recent case studies from successful medicinal chemistry efforts from recent publications to demonstrate how these approaches can be used to move a medicinal chemistry project forward using scaffold hopping techniques.

## Acknowledgments

N.B. is funded by Cancer Research UK Grant No. C309/A8274.

## References


1 Brown, N. and Jacoby, E. (2006) On scaffolds and hopping in medicinal chemistry. *Mini Reviews in Medicinal Chemistry*, **6**, 1217–1229.

2 Langdon, S.R., Ertl, P., and Brown, N. (2010) Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Molecular Informatics*, **29**, 366–385.

3 Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angewandte Chemie, International Edition*, **38**, 2894–2896.

4 Stanton, D.T., Morris, T.W., Roychoudhury, S., and Parker, C.N. (1999) Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *Journal of Chemical Information and Computer Sciences*, **39**, 21–27.

5 Bohl, M., Dunbar, J., Gifford, E.M., Heritage, T., Wild, D.J., Willett, P., and Wilton, D.J. (2002) Scaffold searching: automated identification of similar ring systems for the design of combinatorial libraries. *Quantitative Structure–Activity Relationships*, **21**, 590–597.

6 Böhm, H.-J., Flohr, A., and Stahl, M. (2004) Scaffold hopping. *Drug Discovery Today: Technologies*, **1**, 217–224.

7 Brown, N. (ed.) (2012) *Bioisosteres in Medicinal Chemistry*, Wiley-VCH Verlag GmbH, Weinheim.

8 Koltun, D.O., Vasilevich, N.I., Parkhill, E.Q., Glushkov, A.I., Silbershtein, T.M., Mayboroda, E.I., Boze, M.A., Cole, A.G., Henderson, I., Zautke, N.A., Brunn, S.A., Chu, N., Hao, J., Mollova, N., Leung, K., Chisholm, J.W., and Zablocki, J. (2009) Potent, orally bioavailable, liver-selective stearoyl-CoA desaturase (SCD) inhibitors. *Bioorganic & Medicinal Chemistry Letters*, **19**, 3050–3053.

9 Beaulieu, P.L., Gillard, J., Bykowski, D., Brochu, C., Dansereau, N., Duceppe, J.-S., Haché, B., Jakalain, A., Lagacé, L., LaPlante, S., McKercher, G., Moreau, E., Perreault, S., Stammers, T., Thauvette, L., Warrington, J., and Kukolj, G. (2006) Improved replicon cellular activity of non-nucleoside allosteric inhibitors of HCV NS5B polymerase: from benzimidazole to indole scaffolds. *Bioorganic & Medicinal Chemistry Letters*, **16**, 4987–4993.

10 Bovens, S., Kaptur, M., Elfinghoff, A.S., and Lehr, M. (2009) 1-(5-Carboxyindol-1-yl) propan-2-ones as inhibitors of human cytosolic phospholipase A2α: synthesis and properties of bioisosteric benzimidazole, benzotriazole and indazole analogues. *Bioorganic & Medicinal Chemistry Letters*, **19**, 2107–2111.

11 Trimble, L.A., Chauret, N., Silva, J.M., Nicoll-Griffith, D.A., Li, C.-S., and Yergey, J.A. (1997) Characterization of the *in vitro* oxidative metabolites of the COX-2 selective inhibitor L-766,112. *Bioorganic & Medicinal Chemistry Letters*, **7**, 53–56.

**12** Roy, P., Leblanc, Y., Ball, R.G., Birdeau, C., Chan, C.C., Chauret, N., Cromlish, W., Ethier, D., Gauthier, J.Y., Gordon, R., Greig, G., Guay, J., Kargman, S., Lau, C.K., O'Neill, G., Silva, J., Thérien, M., vanStaden, C., Wong, E., Xu, L., and Prasit, P. (1997) A new series of selective COX-2 inhibitors: 5,6-diarylthiazolo[3,2-*b*][1,2,4]triazoles. *Bioorganic & Medicinal Chemistry Letters*, **7**, 57–62.

**13** Markush, E.A. (1924) Pyrazolone dye and process of making the same. U.S. Patent 1,506,316, August 26.

**14** Brough, P.A., Aherne, A., Barril, X., Borgognoni, J., Boxall, K., Cansfield, J.E., Cheung, K.-M.J., Collins, I., Davies, N.G.M., Drysdale, M.J., Dymock, B., Eccles, S.A., Finich, H., Fink, A., Hayes, A., Howes, R., Hubbard, R.E., James, K., Jordan, A.M., Lockie, A., Martins, V., Massey, A., Matthews, T.P., McDonald, E., Northfield, C.J., Pearl, L.H., Prodromou, C., Ray, S., Raynaud, F.I., Roughley, S.D., Sharp, S.Y., Surgenor, A., Walmsley, D.L., Webb, P., Wood, M., Workman, P., and Wright, L. (2008) 4, 5-Diarylisoxazole Hsp90 chaperone inhibitors: potential therapeutic agents for the treatment of cancer. *Journal of Medicinal Chemistry*, **51**, 196–218.

**15** Drysdale, M.J., Dymock, B.M., Finch, B., Webb, P., McDonald, E., James, K.E., Cheung, K., and Matthews, T. (2006) Isoxazole compounds as inhibitors of heat shock proteins. U.S. Patent 2006/0241106 A1.

**16** Reich, H.J. and Cram, D.J. (1969) Macro rings: XXXVII. Multiple electrophilic substitution reactions of [2.2] paracyclophanes and interconversions of polysubstituted derivatives. *Journal of the American Chemical Society*, **91**, 3527–3533.

**17** Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs: 1. Molecular frameworks. *Journal of Medicinal Chemistry*, **39**, 2887–2893.

**18** Lipkus, A.H., Yuan, Q., Lucas, K.A., Funk, S.A., Bartelt, W.F., III, Schenk, R.J., and Trippe, A.J. (2008) Structural diversity of organic chemistry: a scaffold analysis of the CAS Registry. *The Journal of Organic Chemistry*, **73**, 4443–4451.

**19** Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M., and Waldmann, H. (2007) The scaffold tree: visualization of the scaffold universe by hierarchical scaffold classification. *Journal of Chemical Information and Modeling*, **47**, 47–58.

**20** Langdon, S.R., Brown, N., and Blagg, J. (2011) Scaffold diversity of exemplified medicinal chemistry space. *Journal of Medicinal Chemistry*, **51**, 2174–2185.

**21** Varin, T., Schuffenhauer, A., Ertl, P., and Renner, S. (2011) Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *Journal of Chemical Information and Modeling*, **51**, 1528–1538.

**22** Raymond, J., Gardiner, E., and Willett, P. (2002) RASCAL: calculation of graph similarity using maximum common edge subgraphs. *Computer Journal*, **45**, 631–644.

**23** Nicolaou, C.A., Tamura, S.Y., Kelley, B.P., Bassett, S.I., and Nutt, R.F. (2002) Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *Journal of Chemical Information and Computer Sciences*, **42**, 1069–1079.

**24** Clark, A.M. and Labute, P. (2009) Detection and assignment of common scaffolds in project databases of lead molecules. *Journal of Medicinal Chemistry*, **52**, 469–483.

**25** Welsch, M.E., Snyder, S.A., and Stockwell, B.R. (2010) Privileged scaffolds for library design and drug discovery. *Current Opinion in Chemical Biology*, **14**, 1–15.

**26** Evans, B.E., Rittle, K.E., Bock, M.G., DiPardo, R.M., Fredinger, R.M., Whitter, W.L., Lundell, G.F., Veber, D.F., Anderson, P.S., Chang, R.S.L., Lotti, V.J., Cerino, D.J., Chen, T.B., Kling, P.J., Kunkel, K.A., Springer, J.P., and Hirshfield, J. (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *Journal of Medicinal Chemistry*, **31**, 2235–2246.