Synthetic library design

Christoph M. Huwe

Schering AG, Medicinal Chemistry, Corporate Research, 13342 Berlin, Germany

Compound libraries have an important role in the drug discovery process. Various computational methods are available as decision support tools for medicinal chemists involved in compound library synthesis programs. These methods can be used to assemble a flexible library design scheme consisting of a structure-based library design followed by property-biased library refinement and final selection according to structure–activity-relationship considerations.

Whereas the later stages of lead optimization are typically addressed by single synthesis or manual parallel synthesis of smaller compound series, compound libraries have an important role for the hit identification, hit-to-lead and earlier lead optimization stages of a drug discovery program. The concept of combinatorial chemistry [1,2], in conjunction with automated parallel synthesis, has made a large number of novel compounds synthetically accessible, and progress in synthesis [3,4] and purification [5,6] automation technology has left few limitations on the automated synthesis of compounds. However, compound sample quality standards requiring purification and resource limitations including budget, capacity and timeline issues require selection of a subset of compounds to be synthesized. Moreover, the need to supply compounds with appropriate potency, selectivity, solubility, bioavailability, and so on, which are important reasons for attrition [7,8], in addition to intellectual property and synthesis economics considerations, requires careful selection of candidate structures. A flexible, straightforward synthetic library design strategy is needed to enable medicinal chemists to reach these various goals (Figure 1).

Structure-based library design

To guide the synthetic efforts of a library synthesis program and thus increase the chances of hitting the target, a combination of computational chemistry and combinatorial chemistry can be employed [9,10]. Molecular modelling based on protein cocrystal structures or reliable homology models, in combination with knowledge about the binding mode of a compound class, can provide an excellent starting point for compound synthesis. If no protein structures or homology models are available, design of compounds based on pharmacophore models or analogue design around known hits or privileged structural elements [11,12] can be applied to provide an initial structure concept. The uncertainty of the model can then be addressed by designing a compound library around the structure concept – that is, moving around the groups attached to a scaffold by modifying distances, angles, substitution pattern, and so on. In this way, the potentials of both computational chemistry and combinatorial chemistry can be exploited while simultaneously addressing the weaknesses of each approach. Based on the initial structure concept, building blocks are then preselected by criteria such as synthetic utility, undesired structural elements [13], availability and cost to define an initial virtual library (Figure 2).

As the next step, virtual screening [14] can be used to eliminate building blocks with a low probability of fitting into the binding pocket. In this approach, the affinity of a potential ligand is estimated by how well its structure complements the 3D structure of the target binding site. Although simple docking approaches typically do not enable candidate structures to be ranked, recently a 'hierarchical' virtual screening approach has been suggested to provide more accurate results [15]. This approach starts with a coarse-grain conformational search over a large number of configurations filtered with a fast but crude energy function. This is followed by a succession of finer-grain levels, using successively more accurate descriptions of the ligand–protein–solvent interactions to filter successively fewer cases, and, finally, optimizes one

Corresponding author: Huwe, C.M. (christoph.huwe@schering.de)

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.06.017

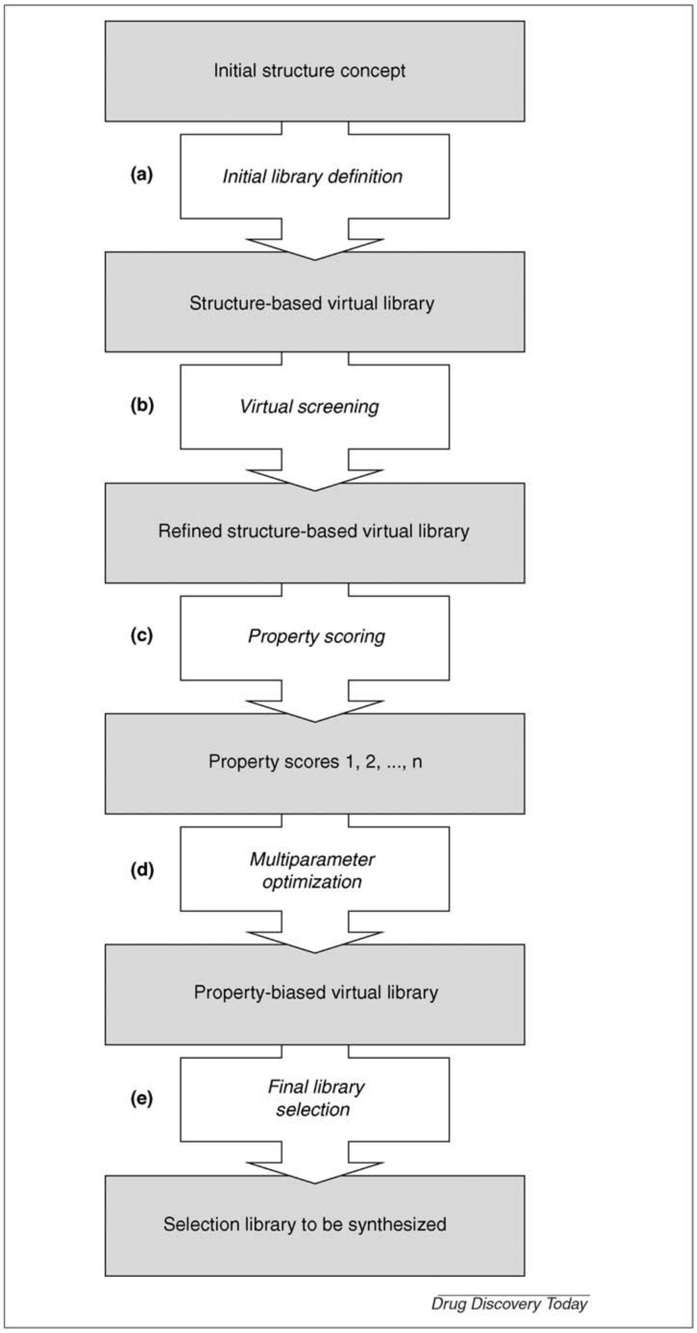
Initial structure concept
(a)
Initial library definition
Structure-based virtual library
(b)
Virtual screening
Refined structure-based virtual library
(c)
Property scoring
Property scores 1, 2, ..., n
(d)
Multiparameter optimization
Property-biased virtual library
(e)
Final library selection
Selection library to be synthesized
Drug Discovery Today

FIGURE 1

Example of a flexible synthetic library design scheme based on the methods discussed in this article. The scheme can be adapted, depending on the stage of the drug discovery process and the specific library goals. **(a)** An initial structure concept is generated by molecular modelling based on protein cocrystal structures or reliable homology models. Using a set of potential building blocks preselected by criteria such as synthetic utility, undesired structural elements, availability and cost, an initial structure-based virtual library is defined. **(b)** Building blocks that are too large to fit into the binding pocket when attached to the scaffold can optionally be eliminated by virtual screening to give a refined structure-based virtual library. **(c)** Calculation of properties to describe the druglikeness and bioavailability of a compound gives several individual property scores. **(d)** Depending on the number of individual scores, a consensus score for each structure can optionally be generated by multiparameter optimization to give a property-biased virtual library. **(e)** Final selection based on structure–activity-relationship considerations and experience gives a selection library with reduced size and enriched desired properties to be synthesized.

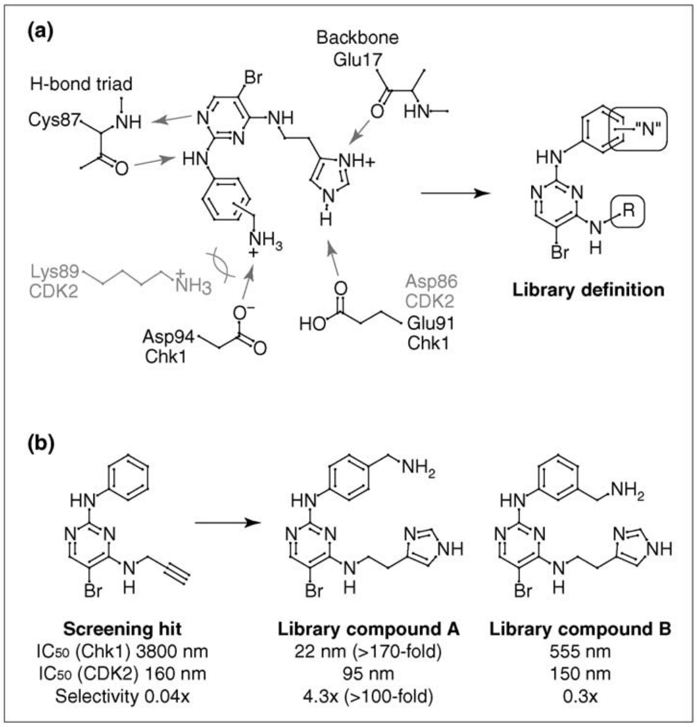

FIGURE 2

Example of a structure-based library design approach in the context of a hit-to-lead project, aiming at the identification of inhibitors of checkpoint kinase 1 (Chk1) showing selectivity towards cyclin dependent kinase 2 (CDK2). (a) Based on a homology model a mutation within Chk1 (Asp94) compared with CDK2 (Lys89) was identified as an opportunity to generate selectivity by adding a positively charged amine group to the phenyl ring (NH_3^+), which should result in a binding interaction with negatively charged Asp94 within Chk1, but a repulsive interaction with positively charged Lys89 within CDK2. In addition, a histamine in the pyrimidine-4-position should help to further increase binding affinity. Based on this concept, a small library was designed, bearing different groups (NHR) attached to the 4-position and, in particular, different amine groups ("N") attached to the aminophenyl in the 2-position. **(b)** One member of the library (library compound A), a 22 nM inhibitor of Chk1, showed significantly improved potency (>170-fold) and moderate Chk1 selectivity (~4-fold, >100-fold improvement) compared with an initial, strongly CDK2-selective screening hit, thus turning around the selectivity within this structure class. A very similar compound (library compound B), however, was still CDK2 selective and significantly less potent. These results demonstrate the power of a structure-based approach, but also illustrate the need to synthesize a library of closely related analogues around a structure concept (instead of single compounds) for maximum impact.

configuration of the ligand in the protein site, using a relatively accurate energy expression and description of the solvent, which would be impractical for all conformations.

Property-biased library refinement

The initial virtual library can then be subjected to a property-biased refinement step. For this purpose, several properties can be calculated that might be used to predict the leadlike or druglike behaviour of members of a virtual library. Generally, product-based design approaches might be preferred over simpler reactant-based designs because they have been demonstrated to give superior results [16].

According to Lipinski *et al.* [17,18], compounds undergoing passive transport that meet two or more of the following criteria tend to have low oral absorption: molecular weight (MW)

>500 Da, number of H-bond donors (expressed by the number of NH + OH groups) >5, number of H-bond acceptors (expressed by the number of N + O atoms) >10 (2 × 5), calculated log of the partition coefficient between *n*-octanol and water (CLogP) >5. This Rule of Five might be extended by including the number of rotatable bonds (RTB) >8 and number of fused rings >5 [19].

Customization of these general rules can be useful, depending on the library goal; for example, more leadlike lead finding libraries or more druglike lead optimization libraries. Leads and oral drugs tend to have similar properties (lower MW and CLogP; smaller number of H-bond acceptors, H-bond donors and RTB) but 'pure' leads with low MW and rather simple structure are often difficult to identify via high-throughput screening (at least at the 10 μM concentration often employed). Because the molecular weight typically increases by ~20–50 Da during a lead optimization program [20], a 'druglike lead' approach using slightly more stringent margins might be a good compromise for most purposes – for example, MW <450 Da, CLogP ≤4.5, H-bond donors ≤5, H-bond acceptors ≤8 [21].

The aqueous solubility of a compound is an important property that influences both the bioavailability as well as the magnitude of many absorption, distribution, metabolism and excretion (ADME) properties, and ultimately the bioactivity of an oral compound. However, solubilization is a complicated process influenced by such factors as lipophilicity, H-bond formation, crystal packing and counterion. Several computational models have been developed to estimate the aqueous solubility from the structure of a compound [22]. Improved results can often be obtained if consensus models are used but the available models are typically only able to make predictions with one log unit uncertainty, and significant progress is only expected from utilization of a large, diverse, standardized set of solubility data of druglike molecules. Therefore, currently the application of a classification model (e.g. 'low' <10 mg/l, 'medium' 10–100 mg/l, 'high' >100 mg/l) might be more useful than using absolute values.

Another property that can add significant value to the library design process is the polar surface area (PSA) of a structure, which can be used to predict Caco-2 permeability (experimental intestinal absorption estimation model). PSA is defined as the sum of the (preferably solvent-accessible) surface contributions of polar atoms, such as O, N (sometimes also S, P) and the attached H atoms, in a molecule. The most accurate methods to calculate this value, dynamic PSA and static PSA, are based on a set of low-energy conformers and a single low-energy conformer, respectively, and require minimization of 3D structures, making these methods too slow to be useful for larger numbers of candidate structures. By contrast, topological 2D fragment-based PSA calculation (TPSA) is fast, correlates reasonably well with 3D PSA and thus is the method of choice for synthetic library design purposes [23]. Statistical analysis has shown that the PSA distribution for orally administered non-CNS drugs is 10–140 Å^2, with a maximum at 50–70 Å^2. Orally administered CNS drugs tend to have smaller PSA values of 0–90 Å^2, with a maximum at 30–50 Å^2 [24]. Additional studies have confirmed the correlation of RTB and PSA with oral bioavailability [25,26].

However, a recent study [27] suggests that the appropriate method for predicting bioavailability strongly depends on the charge of the molecule. According to this analysis, the Rule of Five, but not PSA, is predictive for neutral, zwitterionic and positively charged compounds (at pH 6–7). However, for negatively charged compounds, PSA alone, but not in combination with MW or CLogP, correlates well with bioavailability. These results are summarized in a bioavailability score (ABS), which represents the

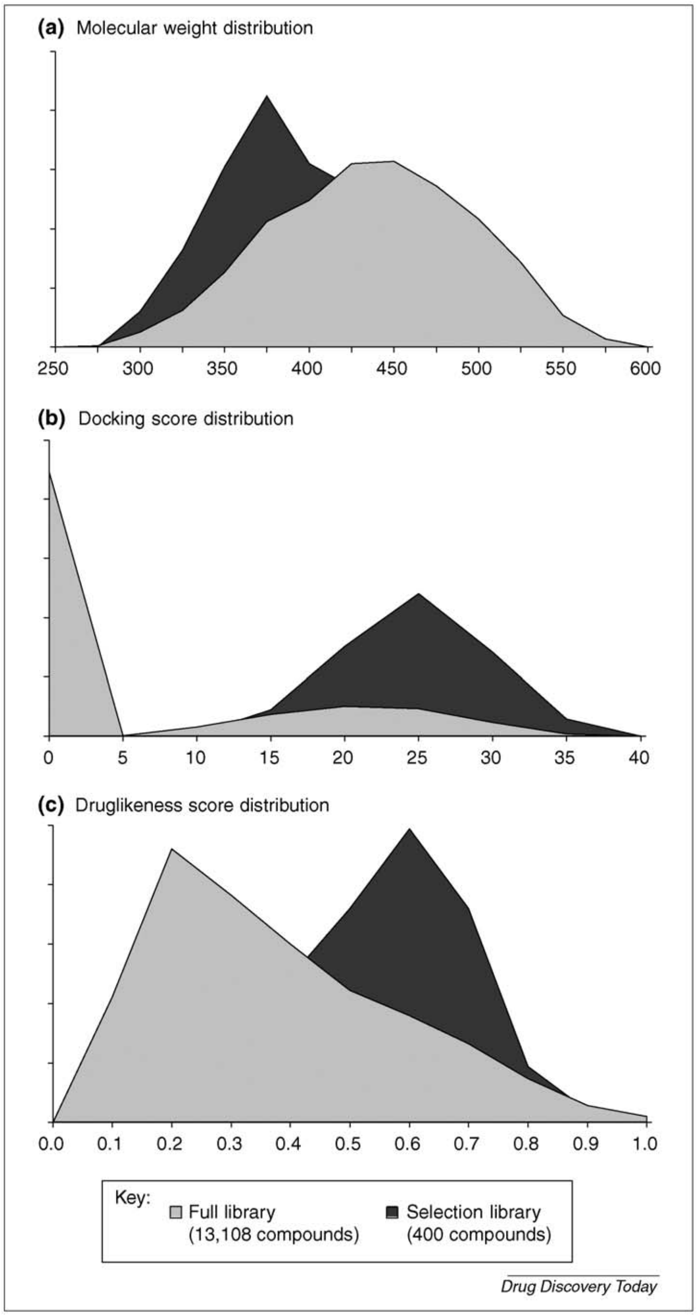

FIGURE 3

Example of the result of a property-biased library refinement process. For a structure-based virtual library of ~13,000 members the molecular weight, a docking score and a druglikeness score were calculated and the results were subjected to a multi-parameter optimization process. In this way a subset library of 400 structures was selected showing **(a)** a molecular weight distribution shifted towards lower weights useful for early optimization efforts, **(b)** a docking score distribution showing enrichment of higher docking scores and complete loss of very low docking scores and **(c)** a druglikeness score distribution shifted towards higher druglikeness. These results illustrate the strong influence of building block selection on the property distribution of a compound library.

probability of a compound of having $>10\%$ bioavailability in the rat or measurable Caco-2 permeability (ABS has also been demonstrated to be related to human bioavailability). ABS is 0.11 for anions for which PSA is >150 Å^2, 0.56 for PSA 75–150 Å^2 and 0.85 for PSA <75 Å^2. For other compounds, ABS is 0.55 if it passes the Rule of Five and 0.17 if it fails. More complicated systems to evaluate druglikeness, using up to 25 parameters, have also been suggested [28].

Likeness models and diversity analysis

Likeness models are used to classify structures based on molecular descriptors, such as MW, RTB, LogP, PSA or fragment-based Ghose–Crippen descriptors [29–31], and machine-learning techniques using a training set. Examples include druglikeness [32,33] but more recently also kinase inhibitor-likeness [34]. Although limitations, including the use of 2D fragment descriptors and the quality of the training set, might limit the predictivity, likeness models are useful tools for selecting subsets from large numbers of structures for target family-oriented focused libraries.

Despite being an important tool in corporate compound collection design, molecular diversity analysis approaches [35], which enable the calculation of the 'unrelatedness' of structures, have limited utility in synthetic library design. However, these approaches can be applied to reduce the number of candidate structures significantly when dealing with large initial virtual libraries ($>10^5$–10^6), for which some of the property calculations described above might be too slow to be useful. The property-biased refinement step can then be applied to the preselected subset. In addition, similar to likeness models, diversity (or, in this case, similarity) analysis might also enable 'scaffold hopping' – that is, finding compounds that have similar biological activity although belonging to different structural classes.

Models to estimate the cytochrome P450 enzyme inhibition potential [36,37], hERG (human ether-a-go-go-related-gene) channel inhibition [38,39] and other *in silico* toxicology predictions [40], as well as metabolic stability predictions [41,42], have also been described but might be more useful in the later stages of compound discovery.

Final compound selection

Based on the property scores generated by the individual models, multiparameter optimization approaches can be used to select candidate compound subsets while simultaneously optimizing several parameters [43]. Using the results of this preselection, the final selection should then be made by the medicinal chemist based on structure–activity-relationship considerations and experience. Although this step is sometimes viewed suspiciously, it provides an important 'reality check' and can also address issues beyond the scope of computational models (Figure 3).

Conclusion

Compound libraries have an important role in the hit identification, hit-to-lead and earlier lead optimization phases of the drug discovery process. Considering the large number of possible structures accessible via modern automated synthesis and purification technologies, the selection of compounds to be synthesized for a program is crucial. In addition to requirements such as building block availability, synthesis economics and intellectual property considerations, the selected compounds need to show appropriate druglikeness, solubility and bioavailability. Various computational models have been developed to enable the classification of virtual compounds according to these requirements before synthesis. The computational methods reviewed herein should therefore be considered as important decision support tools for compound library synthesis programs within medicinal chemistry groups.

Acknowledgements

The author would like to thank the Schering AG computational chemistry, structural biology and automated medicinal chemistry group members for stimulating discussions, particularly Hans Briem, Judith Günther, Anton ter Laak, Hans-Peter Wrona-Metzinger, Emil Eckle, Martina Schäfer and Andreas Steinmeyer.

References

1 Balkenhohl, F. *et al.* (1996) Combinatorial synthesis of small organic molecules. *Angew. Chem. Int. Ed. Engl.* 35, 2289–2337

2 Dolle, R.E. (2005) Comprehensive survey of combinatorial library synthesis: 2004. *J. Comb. Chem.* 7, 739–798

3 Brändli, C. *et al.* (2003) Automated equipment for high-throughput experimentation. *Chimia (Aarau)* 57, 284–289

4 Reader, J.C. (2004) Automation in medicinal chemistry. *Curr. Top. Med. Chem.* 4, 671–686

5 Ripka, W.C. *et al.* (2001) High-throughput purification of compound libraries. *Drug Discov. Today* 6, 471–477

6 Koppitz, M. *et al.* (2005) Maximizing automation in LC/MS high-throughput analysis and purification. *J. Comb. Chem.* 7, 714–720

7 Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discov. Today* 2, 436–444

8 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715

9 Böhm, H.-J. and Stahl, M. (2000) Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* 4, 283–286

10 Leach, A.R. *et al.* (2000) Synergy between combinatorial chemistry and *de novo* design. *J. Mol. Graph. Model.* 18, 358–367

11 Müller, G. (2003) Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* 8, 681–691

12 Bondensgaard, K. *et al.* (2004) Recognition of privileged structures by G-protein coupled receptors. *J. Med. Chem.* 47, 888–899

13 Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 8, 86–96

14 Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* 432, 862–865

15 Floriano, W.B. *et al.* (2004) HierVLS: hierarchical docking protocol for virtual ligand screening of large molecule databases. *J. Med. Chem.* 47, 56–71

16 Green, D.V.S. and Pickett, S.D. (2004) Methods for library design and optimization. *Mini Rev. Med. Chem.* 4, 1067–1076

17 Lipinski, C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249

18 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. *Adv. Drug Deliv. Rev.* 46, 3–25

19 Darvas, F. *et al.* (2002) In silico and ex silico ADME approaches for drug discovery. *Curr. Top. Med. Chem.* 2, 1287–1304

20 Lajiness, M.S. *et al.* (2004) Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Devel.* 7, 470–477

21 Oprea, T.I. (2002) Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput. Aided Mol. Des.* 16, 325–334

22 Jorgensen, W.L. and Duffy, E.M. (2002) Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* 54, 355–366

23 Ertl, P. *et al.* (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* 43, 3714–3717

24 Kelder, J. *et al.* (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* 16, 1514–1519

25 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623

26 Lu, J.J. *et al.* (2004) Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J. Med. Chem.* 47, 6104–6107

27 Martin, Y.C. (2005) A bioavailability score. *J. Med. Chem.* 48, 3164–3170

28 Xu, J. and Stevenson, J. (2000) Drug-like index: a new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* 40, 1177–1187

29 Ghose, A.K. and Crippen, G.M. (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. 1. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* 7, 565–577

30 Ghose, A.K. and Crippen, G.M. (1987) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Comput. Chem.* 8, 21–35

31 Ghose, A.K. *et al.* (1988) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. 3. Modeling hydrophobic interactions. *J. Comput. Chem.* 9, 80–90

32 Ajay, A. *et al.* (1998) Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* 41, 3314–3324

33 Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 41, 3325–3329

34 Briem, H. and Günther, J. (2005) Classifying 'kinase inhibitor-likeness' by using machine learning methods. *Chembiochem* 6, 558–566

35 Perez, J.J. (2005) Managing molecular diversity. *Chem. Soc. Rev.* 34, 143–152

36 Kriegl, J.M. *et al.* (2005) Prediction of human cytochrome P450 inhibition using support vector machines. *Comb. Sci.* 24, 491–502

37 Hutzler, J.M. *et al.* (2005) Predicting drug–drug interactions in drug discovery: where are we now and where are we going? *Curr. Opin. Drug Discov. Devel.* 8, 51–58

38 Cianchetta, G. *et al.* (2005) Predictive models for hERG potassium channel blockers. *Bioorg. Med. Chem. Lett.* 15, 3637–3642

39 Sanguinetti, M.C. and Mitcheson, J.S. (2005) Predicting drug–hERG channel interactions that cause acquired long QT syndrome. *Trends Pharmacol. Sci.* 26, 119–124

40 Vedani, A. *et al.* (2005) *In silico* prediction of harmful effects triggered by drugs and chemicals. *Toxicol. Appl. Pharmacol.* 207, S398–S407

41 Riley, R.J. *et al.* (2005) A unified model for predicting human hepatic, metabolic clearance from *in vitro* intrinsic clearance data in hepatocytes and microsomes. *Drug Metab. Dispos.* 33, 1304–1311

42 Korzekwa, K.R. *et al.* (2004) Models for cytochrome P450-mediated metabolism. *Biotechnology: Pharmaceutical Aspects* 1, 69–80

43 Green, D.V.S. and Pickett, S.D. (2004) Methods for library design and optimization. *Mini-Reviews Med. Chem.* 4, 1067–1076