ELSEVIER



# Methods for mining HTS data

**Gavin Harper and Stephen D. Pickett**

GSK, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, United Kingdom

**Data mining is a fast-growing field that is finding application across a wide range of industries. HTS is a crucial part of the drug discovery process at most large pharmaceutical companies. Accurate analysis of HTS data is, therefore, vital to drug discovery. Given the large quantity of data generated during an HTS, and the importance of analyzing those data effectively, it is unsurprising that data-mining techniques are now increasingly applied to HTS data analysis. Taking a broad view of both the HTS process and the data-mining process, we review recent literature that describes the application of data-mining techniques to HTS data.**

In 2003, a survey of screening methods in 53 laboratories suggested that approaching half of all HTS do not generate any leads at all [1]. This is in spite of continuing increases in the capacity of HTS, with many large companies screening hundreds of thousands to millions of compounds. The success or failure of an HTS is dependent on several factors. These factors include the rational design of the screening collection (which compounds are actually screened), the correct identification of genuine hits, from the often noisy assay data, for further testing (typically only ~50% of apparent hits in primary HTS assays are confirmed to be active in further testing [2]), and the effective prioritization of confirmed hits for follow-up testing or optimization work. There is a clear need for methods that can aid decision making at these crucial decision points.

Data mining has been described as 'the exploration and analysis, by automatic or semiautomatic means, of large quantities of data to discover meaningful patterns and rules' [3], and it finds application across a wide selection of industries. It tends to have developed furthest in industries such as marketing, banking and e-commerce – as a result of these industries typically having large databases of relevant information that is readily available. More recently, there has been an increase in the application of data-mining techniques to bioinformatics, and related disciplines [4]. Given the large quantity of data generated during an HTS, and the importance of analyzing those data effectively, it is unsurprising that data-mining techniques are now increasingly applied to HTS data analysis.

*Corresponding author:* Harper, G. (gavin.x.harper@gsk.com)

## The data-mining process

Various attempts have been made within the data-mining community to analyze the procedure of data mining – in terms of standard processes that can be applied independently of the field of application and the specific algorithms used for model building. Broadly, the target has been to formalize a list of tasks that, taken together, encapsulate the data-mining process. Perhaps the best-known such approach is the cross-industry standard process for data mining (CRISP-DM) [5]. CRISP-DM has proved popular; a 2002 poll of data-mining practitioners (http://www.kdnuggets.com/polls/2002/methodology.htm) found that 51% described CRISP-DM as the 'main methodology used for data mining'.

CRISP-DM splits the data-mining endeavor into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. One striking feature of this delineation of data mining into phases is that modeling is only seen as one component of the data-mining process. Appreciating the context within which data are generated is important for the business understanding and data understanding steps. Data quality is covered as part of the data understanding step, and data cleaning as part of the data preparation step. Models need to be constantly evaluated for their success. All of these aspects of the general data-mining process are highly relevant to the analysis of HTS data.

At a simple level, CRISP-DM points us towards a range of aspects of HTS that we need to engage with as part of the HTS data-mining process. These include issues of data quality, data manipulation and interpretation, model building, and the evaluation of model success.


1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2006.06.006

## Impact of assay quality

The quality of the data available for modeling is highly dependent on the quality and accuracy of the HTS assay. Hence, the use of control wells in screening plates is often augmented by the capture of various quality control parameters that can be used to validate and monitor screening performance [6]. It is becoming increasingly recognized that the monitoring and control of assay performance is crucial to screening success. There are several possible sources of problems that ultimately impact the quality of the assay output. These include effects that are assay specific, technology specific and compound specific. Some examples are discussed in Box 1. Figures 1–3 show examples where the quality of assay output is affected by a particular problem. Pattern recognition techniques can be used to help identify and correct for some of these problems.

In a recent paper, Padmanabha and colleagues [7] discuss various aspects of quality control including assay design, the real-time monitoring of assay performance, sources of error such as incorrect concentrations of samples, and the influence all of these individual aspects have on the success (or otherwise) of subsequent modeling of the data. Woodward *et al.* [8] describe a relatively sophisticated statistical analysis of screening performance using different screening technologies – pointing out some of the pitfalls that occur when using simpler methods of measuring the performance of a screen. Kalos and Rey [9] recently described how, at the Dow Chemical Company, Six Sigma process management was introduced in partnership with data-mining initiatives. As HTS data-mining activities become more common, it seems likely that awareness of data quality issues will rise; and, hence, that some of the quality management and process control methods (more commonly seen in manufacturing environments) will be adopted increasingly by pharmaceutical companies.

## Prioritization of possible actives

HTS commonly involves running a single-shot (single replicate, single compound concentration) assay on a large collection of compounds (known as the primary screen), followed by subsequent follow-up rounds of single-shot and dose–response screening. The number of compounds that can be screened in these subsequent rounds is often limited by budgetary or practical constraints. Thus, the effective prioritization of hits for follow-up work from primary screening is crucial to success [10].

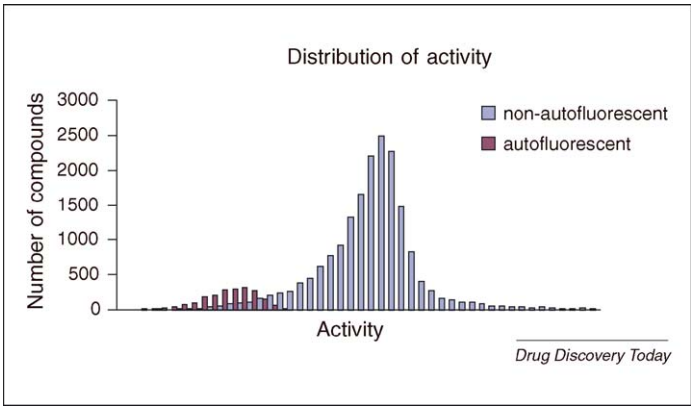


**FIGURE 1**

**Autofluorescent compounds significantly influence the shape of the overall distribution of activities in this set of wells from one screening run of an HTS.**

**BOX 1**

### Common sources of data quality issues within HTS

Inaccuracies in the measurement of activity in HTS can arise for a variety of reasons. False positives are biologically inactive compounds that appear to be active in the primary assay. They cause a particular problem because HTS often has a very low hit rate, so that even when a small percentage of inactive compounds appear as false positives the result can be that the number of false positives exceeds the number of truly active compounds that appear as hits in the screen. With limited resources for following up hits, this can severely affect the overall success of the HTS. False negatives are biologically active compounds that appear to be inactive in the primary assay, and result in interesting compounds being missed in screening. Although false positives are often identified as such in further rounds of screening, the extent of false negatives can be difficult to assess, because no further screening of compounds that were inactive in the primary screen usually takes place.

There are various possible reasons for false positives and false negatives arising in HTS. Often these reasons are technology-specific. The widespread use of fluorescence-based assays can lead to misleading results for fluorescing and quenching compounds [49]. This can give rise to false positives, or extreme negative values (Figure 1) that can impact automated hit-identification methods. Such problems are, perhaps, best dealt with by measuring the autofluorescence of compounds in the appropriate assay medium – utilizing this information for prioritizing hits.

Another problem with more-complex screening formats can arise with compounds that interact via an undesirable mechanism, for example binding to an assay component other than the biological target. A compound-specific source of false positives arises with so-called 'promiscuous inhibitors'. Recent work in this area has been published describing rapid assays for detecting compounds that are acting via a promiscuous mechanism [52]. Neglecting the possible influence of such compounds can lead to misleading correlations, as shown in Figure 2. Appropriate use of detergents in the assay can also help to alleviate this problem [53]. Other sources of compound interference – via covalent modification for example – can be addressed by appropriate substructural filtering of the compound collection, either before screening or at the results annotation stage [54,55].

Other sources of error are the temporal and spatial effects often observed in compound plates [6], and these effects contribute to systematic errors in the measured activity – where the direction and magnitude of the inaccuracy depend on exactly where on the plate (and when during the screening run) a compound is assayed. An example is shown in Figure 3.

Various attempts have been made to correct for some of these systematic errors using, for example, the median polish algorithm [11,56]. There are also several commercial packages available for the identification and correction of systematic errors, however most approaches assume a degree of randomness in the assignment of compounds to plates, so that segments of plates showing a pattern of activity are assumed to be artifacts. Unfortunately, as a result of logistical constraints, structurally similar compounds can often be placed in adjacent wells on the same plate, putting the validity of this assumption in to question.

### *Prioritization of compounds from a 'hit list'*

Identification of a 'hit list' of potentially active compounds from a primary HTS, based on their apparent potency, is a common first

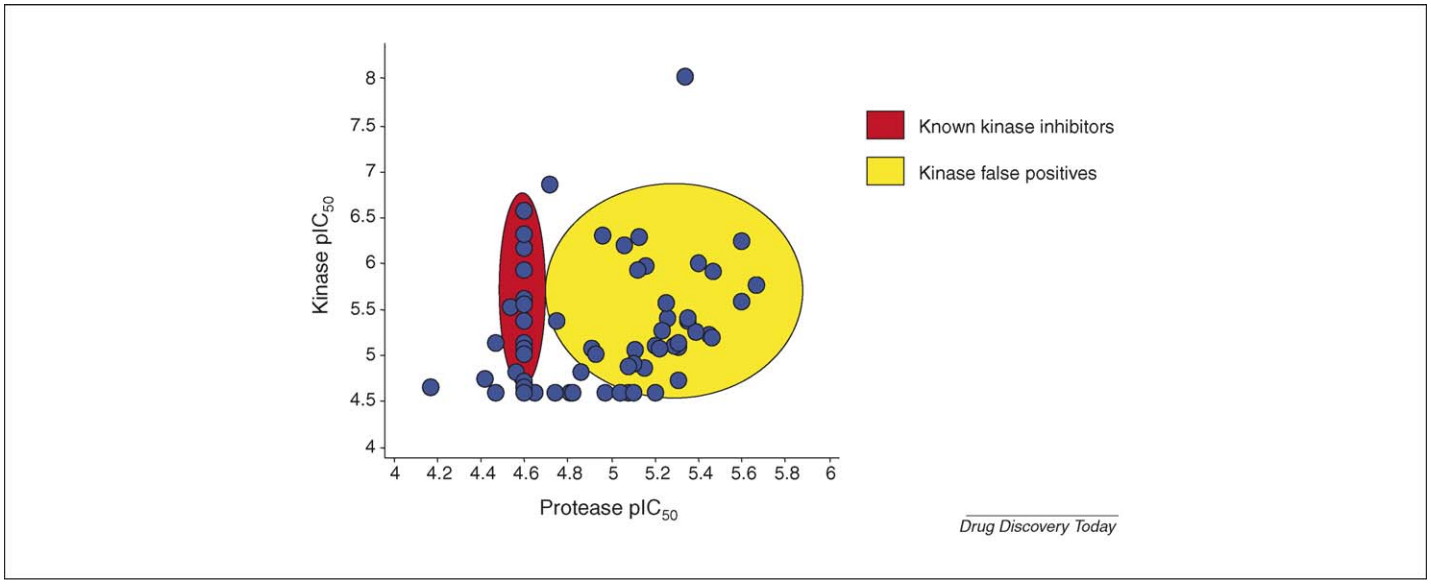


**FIGURE 2**

**An example of promiscuous compounds impacting apparent correlations between screens.** The compounds that appear to be active against the kinase and the protease targets were in fact found to be acting in the kinase screen via a promiscuous mechanism.

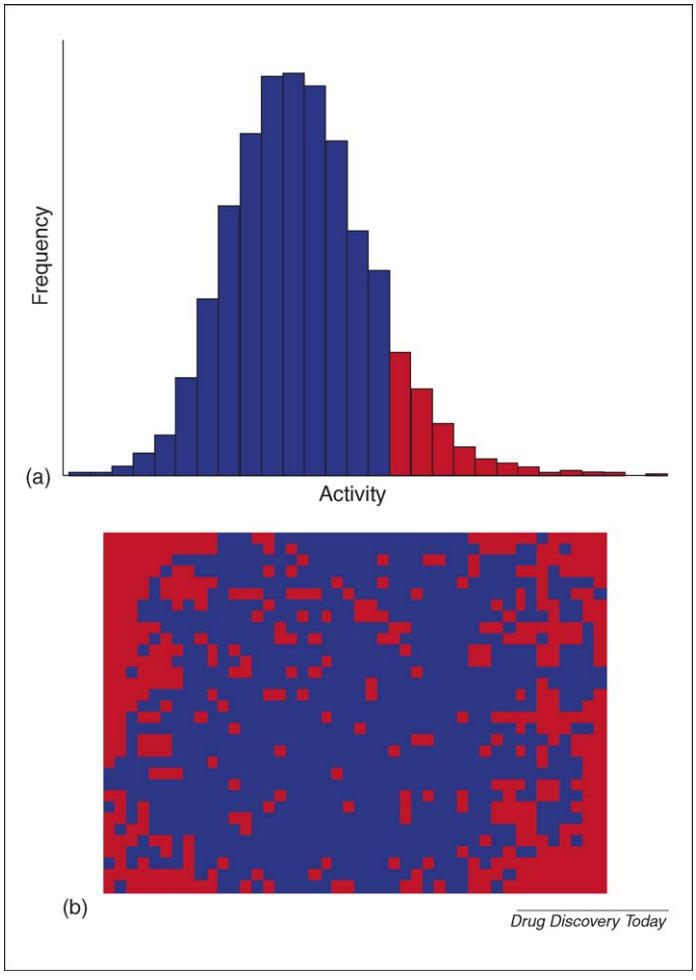


**FIGURE 3**

**A plate displaying edge effects.** The highest measured activities are marked red [in both **(a)** the histogram and **(b)** the plate map]. Looking at the plate map, it is clear that compounds with a high activity are clustered around the edge of the plate in this example.

step in deciding which compounds to test further. Methods based on the distribution of assay readout values, from control wells or the actual samples themselves, are often suggested for defining a threshold assay-readout value. Above such a threshold compounds are deemed likely to be active, and should be included in the hit list [11,12]. These methods should generally be preferred over the use of a predefined activity threshold. Although testing the entire hit list in subsequent rounds of screening is theoretically possible, practical restrictions on screening or compound-handling capacities usually mean that further prioritization of compounds from the hit list is still necessary.

Prioritization of compounds in the hit list is often performed by the expert judgment of a scientist. The scientist personally views the structures, and HTS activities, of the compounds in the hit list, and chooses a subset of the compounds to progress to further testing. The ultimate aim is to ensure that compounds in the subset progressed are the ones that are likely to be active and have drug-like properties, while maintaining good coverage of the chemical diversity of hits from the screen. Given the personal effort required from the scientist in making this compound progression decision, it is particularly important to present the data in a manageable form. One recent article describes PowerMV, which is a package developed to aid decision making in the evaluation of HTS hit lists [13]. The software calculates physical properties thought to be related to drug-likeness and allows near-neighbor searching – based on fragment-based chemical fingerprints. Various groups have previously shown examples of similar procedures, with clustering and interactive filtering (based on calculated or measured properties of compounds) being used to present the data in a form that is easy to browse and process – often using commercial software such as Spotfire for visualization [14]. Interestingly, an assessment of the behavior of medicinal chemists when they choose sets of compounds to progress has exposed poor consistency – when different chemists reviewed the same compounds and even when the same chemist reviewed the same

compounds on two separate occasions [15]. Although a potential cause for concern, this lack of consistency does not in itself imply that the compound selection process does not make a positive contribution relative to selecting compounds at random or purely on a potency basis.

Other groups suggest that, rather than using human intervention, fully automated methods should be used for prioritizing hits. The case for considering ADME–Tox properties early on in the drug discovery process is made by Lipinski and Hopkins [16], criticizing the 'reductionist' approach to drug discovery, embodied by HTS, for its emphasis on potency and selectivity. They argue that (as well as potency) pharmacokinetics, toxicity and biological processes are all relevant to the success of a drug (but are often poorly represented in the original assay). The move towards earlier consideration of ADME–Tox properties (in the drug discovery process) is also discussed in a review article by Bleicher *et al.* [17]. Therefore, it is no surprise that developability criteria, including general concepts such as drug-likeness or lead-likeness [18], have been suggested as being relevant to the fully automated prioritization of hits. Oprea *et al.* [19] present an empirical scheme where compounds are prioritized based on biological potency, testing negative in toxicity-related literature and the possession of good drug-related properties. Other groups have suggested similar schemes [20].

### *Prediction of true actives*

Several modeling techniques exist that go beyond prioritization of hits and actually build models that attempt to predict which compounds are truly active, using all the primary data (rather than restricting consideration to just a hit list of compounds) and descriptors based on the chemical structures of the compounds screened. The range of descriptors that can be used in modeling is extensive and includes molecular fragments, graph indices and 3D descriptors, among others. An extensive review of molecular descriptors is given by Leach and Gillet [21]. Some data preprocessing can also be helpful before commencing modeling work, as suggested (for instance) by Kolossov and Lemon [22].

Recursive partitioning, in particular, is a well-established procedure for modeling HTS data [23,24]. Variants on a naïve Bayes model have recently been proposed [25,26] for modeling screening data, although in practice these are very similar to earlier work that was described as 'substructural analysis' [27]. A recent paper demonstrates that recursive partitioning, a naïve Bayes model and a support-vector machine are all effective at modeling HTS data in the presence of artificially added stochastic noise [28]. Another methodology that uses potency information to highlight interesting groups of compounds is the data-driven clustering methodology [29]. This uses a variety of molecular representations and descriptors (in combination with HTS data) to identify sizeable structural clusters of predominantly active compounds, with each cluster of compounds based on a single structural descriptor.

A recent paper by Yan *et al.* [30] suggests scoring compounds for progression, based on not only the activity of the compound itself but also the activity of the set of compounds sharing the same scaffold. Scaffolds that are predominantly contained in compounds showing high activity in the screen are favored. Other similar, recent work clusters molecules using structural fingerprints, and scores compounds based on the activities of all the compounds in the same cluster [31]. As Yan points out [30], these methods rely on the chemical redundancy present in most screening collections, and are unlikely to progress small clusters (in particular singletons). The corollary to this is that large clusters of compounds with a common scaffold are more likely to progress. This is an inevitable result of the statistical framework adopted. Simply, there is not a large enough sample size with a singleton or small cluster to conclude that the associated scaffold is linked to activity. The approach leads to an improved proportion of the compounds (progressed for further testing) being confirmed active. However, the structural diversity of the compounds selected for activity confirmation is likely to be lower than if a diverse selection of compounds that were sampled from smaller clusters and singletons were progressed – so whether Yan's approach is more successful in reality will depend on the trade-off between a potentially decreased diversity and an improved activity confirmation rate.

Many automated methods of hit selection have the potential problem of undersampling chemical diversity from regions of chemical space that are more sparsely represented in the screening collection. The diversity of hits available from an HTS is, in part, dependent on the diversity of the set of compounds screened, leading to an important aspect of the design of the HTS experiment – the design of the compound screening collection itself. Theoretical work by Harper and colleagues [32] addresses the optimal balance between focus (on areas of chemical space with higher probabilities of activity) and structural diversity in a collection of molecules to be screened. Although this work concentrates on design of the corporate screening collection, the conclusions, in terms of the impact of diversity on screening success, are also applicable to the choice of compounds to progress from a hit list. Recent reviews of other approaches and trends in compound-collection design are given by Lumley [33] and Schuffenhauer *et al.* [34].

The methods previously described use single-shot screening data for the prediction of activity. In that context, any model generated from the data is used to predict the likelihood of activity for compounds that were themselves already present in the HTS. However, many of these methods can also be used to predict the activity of compounds that were not in the original HTS. The compounds searched could be others from the corporate collection, compounds available from external suppliers, or even virtual compounds that have not yet been synthesized. Shanmugasundaram *et al.* [35] describe the use of a sequence of near-neighbor searches using initial hits from HTS as probe compounds, where several different descriptor spaces are used to define the near-neighbors in each search. In each descriptor space, near-neighbors are found to be three to four times more likely to be hits than randomly selected compounds, but with very little overlap between the near-neighbor lists generated using the different descriptors. The approach is sufficiently straightforward as to be completely automated, and could clearly also be used to rescue false negatives from screening. Engels *et al.* [36] describe another potential approach for the automated identification of false negatives – using fragment-based descriptors and logistic regression.

## Recent trends in HTS data mining

The potential of a group of machine learning techniques, known as kernel methods, for modeling screening data are beginning to be recognized. To date, most attention has focused on the support-vector machine [37,38]. Kernel methods are likely to prove to be a

fruitful area for further research, because they allow kernel functions to be defined directly, in terms of the comparison between graph representations for different molecules [39]. A wide variety of conventional multivariate statistical approaches that are used in chemometrics (principal component analysis, partial least squares and canonical correlation analysis) have kernel variants [40], and are already finding applications in bioinformatics [41,42].

Easy accessibility to data will be important for effective data mining in the future, particularly as the emphasis moves to the identification of compounds with defined profiles of activity across a range of targets, rather than activity in one screen [43]. Methodologies such as affinity fingerprints have been developed that utilize this information across multiple screens when comparing compounds [44–46]. Chemical, screening and genomics database systems have not been tightly integrated. Closer integration of these database systems will enable more-complex questions to be asked of HTS data [47,48].

High-content screening (HCS), where there are multiple variables measured per screening well, is becoming increasingly common. HCS data potentially include spatial and temporal information, presenting new challenges for data interpretation and analysis. Analysis of HCS data can also potentially aid in the identification of false positives, because analysis of all the data available regarding a single well can reveal information that is not captured by a single summary value of activity [49]. One example of HCS data analysis is given by a recent paper that describes data interpretation and hit identification of HCS data, gathered using the Beckman-Coulter Q3DM instrument [50]. Fragment-based screening is also becoming increasingly popular. Analysis of molecular fragments potentially introduces new challenges in screening-collection design and data mining, as traditional molecular descriptors become less relevant [51].

## Conclusion

High attrition rates and the need to reduce the time from target identification to bringing a drug to the marketplace are placing unprecedented levels of demand on HTS, to deliver high-quality hits and leads. Increasingly, these demands include not only improvements in the speed and quality of the HTS but also the generation of compounds that have good selectivity and developability profiles. As the demands placed on HTS increase and evolve so do the requirements in terms of data mining. At the present time, most data-mining methodology is focused on finding hits efficiently from an individual HTS. The challenge in this first phase of HTS data mining has chiefly been to analyze large quantities of single-response data. HCS, as well as the need to analyze data from multiple screens to assess selectivity, developability and polypharmacology, makes it likely that the next phase of HTS data mining will focus on the generation and analysis of high-quality multidimensional data.

## References


1 Fox, S. *et al.* (2004) High-throughput screening: searching for higher productivity. *J. Biomol. Screen.* 9, 354–358
2 Karnachi, P.S. and Brown, F.K. (2004) Practical approaches to efficient screening: information-rich screening protocol. *J. Biomol. Screen.* 9, 678–686
3 Berry, M.J.A. and Linoff, G. (1997) *Data mining techniques for marketing, sales, and customer support*. John Wiley and Sons
4 Wang, J.T.L. *et al.* eds (2004) *Data mining in bioinformatics*, Springer
5 Shearer, C. (2002) The CRISP-DM model: the new blueprint for data mining. *Journal of Data-Warehousing* 5, 13–22
6 Zhang, J.-H. *et al.* (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* 4, 67–73
7 Padmanabha, R. *et al.* (2005) HTS quality control and data analysis: a process to maximize information from a high-throughput screen. *Comb. Chem. High Throughput Screen.* 8, 521–527
8 Woodward, P.W. *et al.* (2006) Improving the design and analysis of high-throughput screening technology comparison experiments using statistical modelling. *J. Biomol. Screen.* 11, 5–12
9 Kalos, A. and Rey, T. (2005) Data mining in the chemical industry. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 763–769
10 Alanine, A. *et al.* (2003) Lead generation – enhancing the success of drug discovery by investing in the hit to lead process. *Comb. Chem. High Throughput Screen.* 6, 51–66
11 Gribbon, P. *et al.* (2005) Evaluating real-life high-throughput screening data. *J. Biomol. Screen.* 10, 99–107
12 Fogel, P. *et al.* (2002) The confirmation rate of primary hits: a predictive model. *J. Biomol. Screen.* 7, 175–190
13 Liu, K. *et al.* (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. *J. Chem. Inf. Model.* 45, 515–522
14 Leach, A.R. *et al.* (2001) SIV: a synergistic approach to the analysis of high-throughput screening data. *221st National Meeting of the American Chemical Society*, 1–5 April 2001, San Diego, CA, U. S. A. (Abstract 080-CINF)
15 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896
16 Lipinski, C. and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature* 432, 855–861
17 Bleicher, K.H. *et al.* (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378
18 Leeson, P.D. *et al.* (2004) Drug-like properties: guiding principles for design – or chemical prejudice? *Drug Discov. Today Technol.* 1, 189–195
19 Oprea, T.I. *et al.* (2005) Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J. Biomol. Screen.* 10, 419–426
20 Lajiness, M.S. and Shanmugasundaram, V. (2004) Strategies for the identification and generation of informative compound sets. In *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery* (Bajorath, J., ed.), pp. 111–129, Humana Press
21 Leach, A.R. and Gillet, V.J. (2003) *An Introduction to Chemoinformatics*. Kluwer Academic Publishers pp. 53–76
22 Kolossov, E. and Lemon, A. (2006) Medicinal chemistry tools: making sense of HTS data. *Eur. J. Med. Chem.* 41, 166–175
23 Rusinko, A., III *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026
24 van Rhee, A.M. *et al.* (2001) Retrospective analysis of an experimental high-throughput data set by recursive partitioning. *J. Comb. Chem.* 3, 267–277
25 Bender, A. *et al.* (2004) Molecular similarity searching using atom-environmaents, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* 44, 170–178
26 Rogers, D. *et al.* (2005) Using extended-connectivity fingerprints with Laplacian-modeified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* 10, 682–686
27 Ormerod, A. *et al.* (1989) Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-. Act. Rel.* 8, 115–129
28 Glick, M. *et al.* (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* 46, 193–200
29 Harper, G. *et al.* (2004) The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* 44, 2145–2156
30 Yan, S.F. *et al.* (2005) Novel statistical approach for primary high-throughput screening hit selection. *J. Chem. Inf. Model.* 45, 1784–1790
31 Klekota, J. *et al.* (2005) Identifying biologically active compound classes using phenotypic screening data and sampling statistics. *J. Chem. Inf. Model.* 45, 1824–1836

32 Harper, G. *et al.* (2004) Design of a compound screening collection for use in high throughput screening. *Comb. Chem. High Throughput Screen.* 7, 63–70
33 Lumley, J.A. (2005) Compound selection and filtering in library design. *QSAR Comb. Sci.* 24, 1066–1075
34 Schuffenhauer, A. *et al.* (2004) Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screen.* 7, 771–781
35 Shanmugasundaram, V. *et al.* (2005) Hit-directed near-neighbor searching. *J. Med. Chem.* 48, 240–248
36 Engels, M.F.M. *et al.* (2002) Outlier mining in high throughput screening experiments. *J. Biomol. Screen.* 7, 341–351
37 Burbidge, R. *et al.* (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* 26, 5–14
38 Byvatov, E. *et al.* (2003) Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput Sci.* 43, 1882–1889
39 Mahé, P. *et al.* (2005) Graph kernels for molecular structure-activity relationship analysis with support vector machines. *J. Chem. Inf. Model.* 45, 939–951
40 Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press
41 Eriksson, L. *et al.* (2004) Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.* 380, 419–429
42 Yamanishi, Y. *et al.* (2004) Heterogeneous data comparison and gene selection with Kernel canonical correlation analysis.. In *Kernel Methods in Computational Biology* (Schölkopf *et al.* eds), pp. 209–229, MIT Press
43 Morphy, R. and Rankovic, Z. (2005) Designed multiple ligands. An emerging drug discovery paradigm. *J. Med. Chem.* 48, 6523–6543
44 Comess, K.M. and Schurdak, M.E. (2004) Affinity-based screening techniques for enhancing lead discovery. *Curr. Opin. Drug Discov. Devel.* 7, 411–416
45 Schuffenhauer, A. and Jacoby, E. (2004) Annotating and mining the ligand-target chemogenomics knowledge space. *Drug Discov. Today Biosilico* 2, 190–199
46 Fliri, A.F. *et al.* (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U. S. A.* 102, 261–266
47 Lutz, M.W. *et al.* (2005) Managing genomic and proteomic knowledge. *Drug Discov. Today: Technol.* 2, 197–204
48 Gund, P. *et al.* (2004) Can drug discovery be industrialised? *Curr. Opin. Drug Discov. Devel.* 7, 283–284
49 Gribbon, P. and Sewing, A. (2003) Fluorescence readouts in HTS: no gain without pain? *Drug Discov. Today* 8, 1035–1043
50 Smellie, A. *et al.* (2006) Visualization and interpretation of high content screening data. *J. Chem. Inf. Model.* 46, 201–207
51 Schuffenhauer, A. *et al.* (2005) Library design for fragment based screening. *Curr. Top. Med. Chem.* 5, 751–762
52 Feng, B.Y. *et al.* (2005) High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* 1, 146–148
53 Ryan, A.J. *et al.* (2003) Effect of detergent on ''promiscuous'' inhibitors. *J. Med. Chem.* 46, 3448–3451
54 Rishton, G.M. (1997) Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today* 2, 382–384
55 Jacoby, E. *et al.* (2005) Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* 5, 397–411
56 Brideau, C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* 8, 634–647