ELSEVIER

Computational chemistry-driven decision making in lead generation

Volker Schnecke and Jonas Boström

Computational Lead Discovery, Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden

Novel starting points for drug discovery projects are generally found either by screening large collections of compounds or smaller more-focused libraries. Ideally, hundreds or even thousands of actives are initially found, and these need to be reduced to a handful of promising lead series. In several sequential steps, many actives are dropped and only some are followed up. Computational chemistry tools are used in this context to predict properties, cluster hits, design focused libraries and search for close analogues to explore the potential of hit series. At the end of hit-to-lead, the project must commit to one, or preferably a few, lead series that will be refined during lead optimization and hopefully produce a drug candidate. Striving for the best possible decision is crucial because choosing the wrong series is a costly one-way street.

With the rise of HTS in the 1990s, compound collections in big pharma needed to grow significantly to keep up with the pace of screening technology. In addition to compound acquisition, combinatorial chemistry was the method of choice for producing diverse libraries of compounds that might act as starting points for new leads in coming projects. Although it is nowadays feasible to maintain collections of a million samples and routinely screen them if protein production or assay capacity permit, it is arguable whether this is always necessary. When sufficient knowledge about the target is available, other approaches, such as focused and sequential screening, are often relevant alternatives.

We will describe several computational techniques that are applied at different stages of hit identification and the hit-to-lead process, and outline recent developments and their successful application. Figure 1 provides an overview of the different tasks and computational techniques used during HTS-based lead generation. For focused screening, most of these methods can also be used for selection of subsets. Because our focus is on computational techniques, we would recommend other reviews for a more general description of the lead generation process [1,2].

Corresponding author: Schnecke, V. (volker.schnecke@astrazeneca.com)

Hit identification: finding actives

Different approaches can be applied for finding actives to start a drug discovery project. When the structure of a target protein is available, structure-based lead generation is a very reasonable option [3]. To follow up on a competitor's lead, scaffold hopping can be applied scaffold hopping can be applied, which is the identification of molecules with analogous structure but significantly different backbone [4]. Alternatively, when a series of known actives are available, a pharmacophore model can be built and used for virtual screening [5]. However, if one is interested in finding a larger number of novel starting points there are, in principle, three approaches to discovering actives in the existing compound collection (Figure 2):

- *Random screening* (*or full HTS*) tests all, or a large subset of all available compounds, typically in several steps starting with a high-throughput primary assay, followed by one or more steps to derive hits (i.e. identity- and purity-controlled samples that show low micromolar activity in a concentration-response screen).
- *Focused screening* is screening of a subset of the compound collection that has been chosen based on knowledge about compounds that are active against a specific target. When there are bottlenecks in assay capacity or protein supply, a focused screen

1359-6446/06/$ – see front matter PII: S1359-6446(05)03703-7

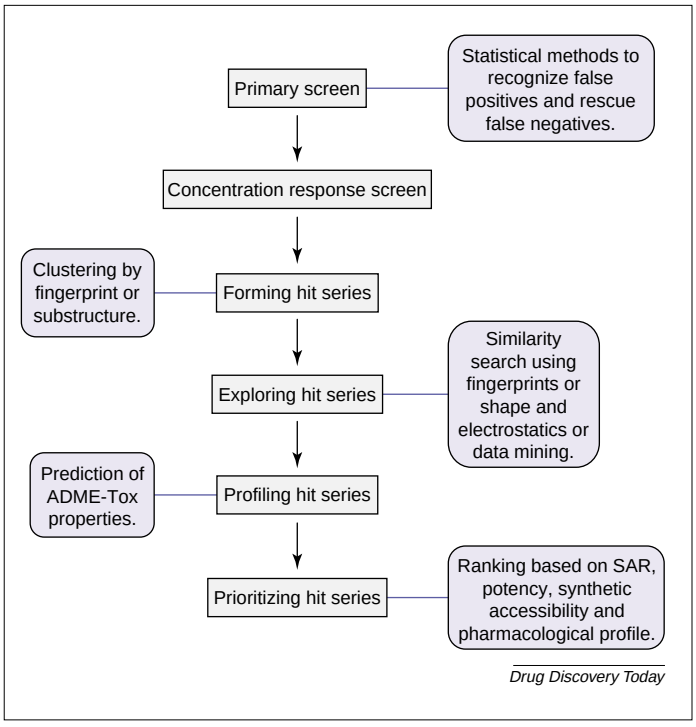


FIGURE 1

An overview of the workflow for an HTS-based lead generation campaign. The diagram shows the computational techniques that are used at the various stages to support data organization, series exploration and prioritization, and guide the overall decision-making process.

is the only way to identify a large number of starting points. It is also the method of choice for finding backup series for a project that has been through a full HTS in the past.

- *Sequential screening* is carried out in several iterations, starting from a representative (diverse) subset of the complete collection. In each iteration, clusters of active compounds are expanded further until a sufficient number of hit series is identified.

Although sequential screening sounds very reasonable from the perspective of a computational chemist, logistics and time lines are not in favour of this approach because the infrastructure is built for large-scale HTS campaigns. Screening the full collection is usually less effort than replating samples and setting up equipment for several independent runs of an assay. However, if rapid picking and reformatting facilities are available, focused or sequential screenings are reasonable alternatives to a full HTS campaign. In addition to lower costs due to lower consumption of protein, substrates and, of course, compound samples, the best argument for focused screening could be that it opens the door for assays that provide a more reliable readout than high-throughput primary screens [6]. When starting with a full HTS, hit identification deals with recognizing false positives, so that hit series need to be dropped when more data become available*. Nevertheless, HTS is still the best method for providing novel starting points, and several success stories of lead compounds that originated from HTS campaigns – including marketed drugs – have been described [7–9].

*One of our co-workers recently renamed hit identification to hit elimination, which unfortunately seems to be a more suitable description of workflow when starting from a full HTS.

Today, big pharma have revised their view of HTS as an all-purpose lead-generation instrument. It is now seen as a toolbox, supporting large-scale HTS and focused medium-throughput screening of targeted libraries, and more reports of successful focused screening campaigns are appearing [10–15]. In order for this flexible use of HTS to be effective, it is crucial that an information-driven decision process is in place to use the available technology optimally and to exploit the generated data [16,17].

Primary HTS data: false positives

When running a full HTS, it is questionable if one should carry out computational work with primary data. Actives in a primary screen can be false positives for several reasons, such as interference with the assay [18] or aggregate formation [19]. However, redundancies are expected in HTS collections, so that statistical methods should be robust enough to analyze primary data. For example, the higher the number of actives that are structurally related to a certain inactive, the higher the probability of it being a false negative. Based on this concept, Engels *et al.* [20] from Janssen Pharmaceuticals use a statistical technique – logistic regression – to rescue false negatives after a primary screen. For dealing with false positives, Schreyer *et al.* [21] have recently proposed an approach that they name data shaving, where features of primary inactives are used to deprioritize follow-up compounds from similarity searches of HTS hits. The noise level in primary data is even higher when mixtures of samples are screened. For this scenario, Glick *et al.* [22] of Novartis suggest another statistical technique – a naïve Bayes classifier – to prioritize compounds for follow-up.

Pharmacopeia prepare their screening libraries using a modified version of split-and-mix combinatorial chemistry, which results in inherent redundancy in their screening experiments that can be statistically utilized to reduce false-positive rates [23]. In their paper, Diller and Hobbs [23] use this data to derive rules for properties that increase the likelihood of biological activity, in analogy to Lipinski's rules for oral bioavailability. They suggest that their filters should be used to increase hit rates of targeted libraries.

When starting with a full HTS campaign, it is best to avoid prioritizing actives until quality data becomes available, that is concentration-response curves for activity, and purity and identity controls of the screening samples. Actives that pass these criteria are then referred to as hits, and from here on there is no difference in hit-to-lead campaigns that originate from focused screening compared to those started with a full HTS (Figure 2).

Clustering: grouping hits

All hits are usually clustered into series to provide an easy overview of the different chemical classes that have been identified as being active against the target. Automated clustering approaches depend on a particular molecular representation and a measure of similarity for a pair of compounds. The most common representations are binary fingerprints, which are used together with the Tanimoto similarity coefficient. Binary fingerprints encode molecular structures in a string of bits (i.e. 0s and 1s) that describe the absence or presence of a certain feature (e.g. a particular functional group). Bit string comparisons are computationally highly efficient, and the Tanimoto coefficient relates the number of set bits (1s) that two fingerprints have in common to the total number of bits set in both fingerprints.

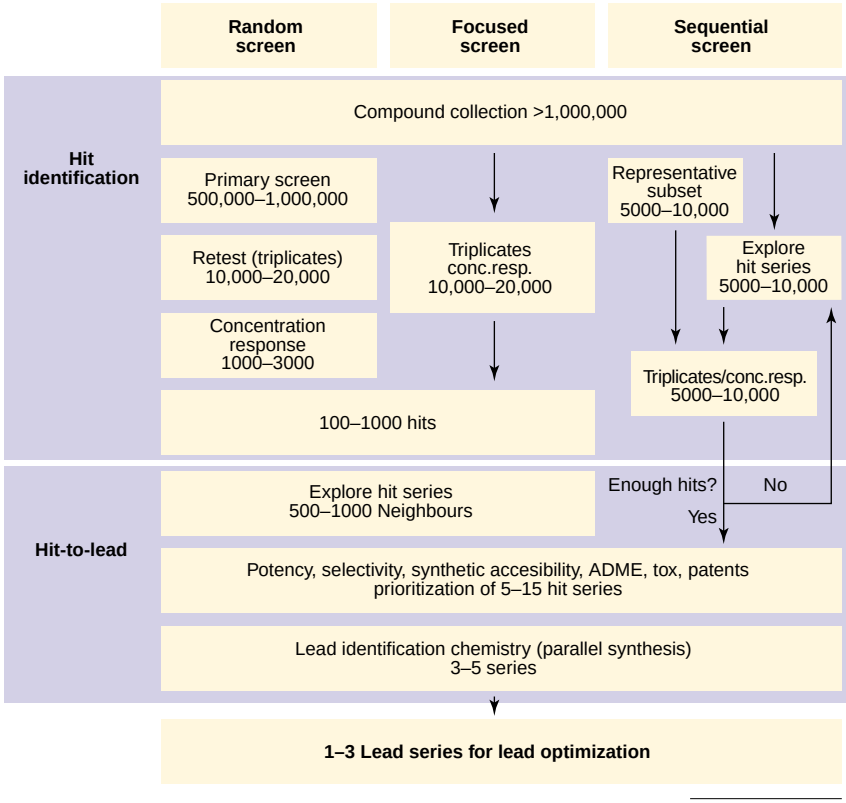


FIGURE 2

A typical view of the overall lead-generation process, which starts with one of three parallel tracks. The numbers are examples, and deviations are expected for certain targets. When series are explored by screening neighbours to interesting hits, these can be taken either from the in-house compound collection or commercial sources.

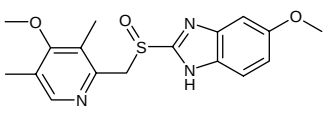
In addition to structural keys, such as ISIS keys from MDL, which encode the presence of functional groups, fingerprints can be derived from paths in the molecular structure, such as Daylight fingerprints, or represent combinations of the two, as implemented in UNITY fingerprints from Tripos. Other implementations of fingerprints are based on 2D or 3D pharmacophores. Here, molecules with matching fingerprints are capable of making similar interactions rather than being similar in structure. All available fingerprints and similarity measures provide different representations of the chemical space, and it needs to be stressed that compounds that are close in one representation will not necessarily be rendered very similar in another one [12].

When a representation of structural features and a similarity coefficient have been decided, there are still different ways of clustering active compounds and arranging them into distinct hit series [24]. The number of compounds to cluster usually dictates the technique. Hierarchical clustering can be either bottom-up (agglomerative), starting from single compounds and joining them into clusters, or top-down (divisive), starting from the whole dataset and splitting it into smaller clusters. Hierarchical clustering methods do not scale to more than several thousand compounds because they require the computation of a similarity measure for each pair of structures. For larger sets, non-hierarchical clustering algorithms, like exclusion sphere, k-nearest neighbours or Jarvis–Patrick, are applicable. A related technique is called partitioning, where the molecules are grouped by subdividing the chemical space. Because this does not involve pairwise similarity comparisons between molecules, there is no limit to the number of compounds that can be grouped by partitioning.

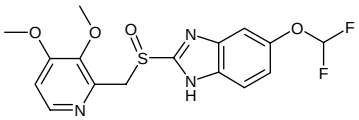
A common drawback of automated clustering methods is that medicinal chemists will always find clusters that they disagree with or identify singletons that should have been added to a certain cluster based on their subjective perception of similarity. Medicinal chemists usually appreciate the presence of common structural features in clustered molecules, which is not always apparent with fingerprint-based clustering. This issue is addressed by approaches that cluster molecules that share a set of substructures, even when these substructures are connected by different linkers [25,26]. An interesting approach has recently been introduced by Stahl and Mauser [27] of Roche, who recommend a two-step process: first, exclusion-sphere clustering based on Tanimoto coefficient for Daylight fingerprints and, second, clustering based on maximum common substructure. In addition to delivering clusters that are more in-line with chemists' expectations, they report that this method is capable of clustering 750,000 compounds.

Exploration: extending hit series

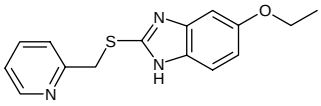
When interesting hits have been identified, there might be a need to search for close analogues to some of them. Especially in the case of singletons or smaller clusters, additional compounds should be screened so that it is possible to derive an initial SAR for the



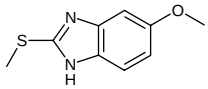
Omeprazole (CNC14367001)



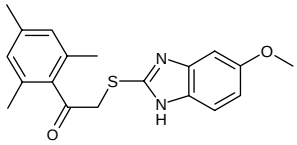
CNC15613753



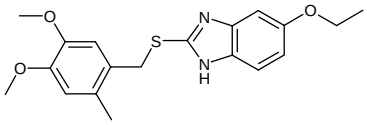
CNC25193116



CNC13721242



CNC18969093



CNC42350322

	Daylight			MDL Keys			Scitegic ECFP6			Rank by Z score
	Tan	Z score	Rank	Tan	Z score	Rank	Tan	Z score	Rank	
CNC14367001	1.00	–18.7	1	1.00	–7.1	1	1.00	–43.9	1	1
CNC15613753	0.83	–14.3	2	0.91	–5.9	3	0.46	–18.5	3	3
CNC25193116	0.79	–13.3	6			>1000	0.20	–6.2	870	10
CNC13721242	0.58	–8.2	207			>1000	0.33	–12.4	10	16
CNC18969093	0.54	–7.1	511	0.75	–3.9	26	0.31	–11.4	30	35
CNC42350322	0.69	–11.0	17	0.75	–4.0	18			>1000	59

Drug Discovery Today

FIGURE 3

Different molecular representations provide different hit lists for fingerprint-based similarity searches. A database of 14 million ChemNavigator compounds was searched for the closest 1000 neighbours to omeprazole using three fingerprint implementations. Merging all 3000 hits and sorting by Z Score instead of Tanimoto (Tan) positions the shown compounds within rank 1 to 59. Even though we do not know about their activity, all of these would be reasonable analogues to screen if omeprazole were an interesting hit out of an HTS campaign that needed follow-up. When depending only on a single fingerprint, some of the hits were not even within the top 1000 neighbours.

corresponding hit series. Different similarity-search methods can be used for this task, and it is recommended that all possible representations are used together, even combining 2D fingerprints with pharmacophore or shape-based searches to exploit the strengths of the individual representations [12,28]. For this, results from independent searches with different representations are done, and the corresponding hit lists are combined using different approaches [29,30].

We have found that the Z score of Tanimoto coefficients are an effective and simple way to combine hit lists from different fingerprint-similarity searches of the same template molecule. The Z score is based on the distribution of the Tanimoto coefficient for all structures that were compared during the similarity search. It indicates how far the Tanimoto coefficient of a compound deviates from the mean Tanimoto coefficient of all compounds, and is expressed in units of standard deviation of the distribution. Tanimoto values depend on the representation being used and cannot be compared across fingerprints. Many approaches described in literature use a general cut-off of 0.85. This was originally proposed for Unity fingerprints, and propagated by several groups, who claimed that compounds above this cut-off are most likely to have similar activity. In a more recent study, Martin *et al.* [31] of Abbott have shown that it is not possible to provide a general cut-off for the Tanimoto coefficient.

Figure 3 shows an example of the advantage over single methods of combining hit lists from different fingerprints. It presents results for the hypothetical case where omeprazol is a member of a cluster that needs to be explored. Five representative hits from a combined search in 14 million ChemNavigator compounds are shown. Although some of them are not picked up within the top 1000 hits of a single fingerprint, they are all found within the top 59 hits when ranked by Z score.

When exploring clusters, it can be an advantage to carry out simultaneous searches with more than one template molecule. When combining hit lists from different templates, the size dependence of a coefficient can bias the results [32,33]. However, it has been shown that using more than a single template is beneficial but that the results depend on the diversity within the set of template molecules [34,35].

Most useful is the combination of different techniques for identifying analogues. Ideally, these methods should be orthogonal in their representation, and a perfect complement to fingerprint or substructure-based similarity search are methods that ignore all structural features of the molecules during comparison, and just look at the overall shape or electrostatic similarity.

Shape and electrostatics: another view on similarity

Several forms of 3D descriptors containing shape and electrostatic information are available today. The underlying idea of shape comparison is that molecules that look similar are likely to act in the same way. The power of shape, as defined by Grant *et al.* [36], is that it is a fundamental molecular property and that shape difference forms a metric space. There are no arbitrary parameters

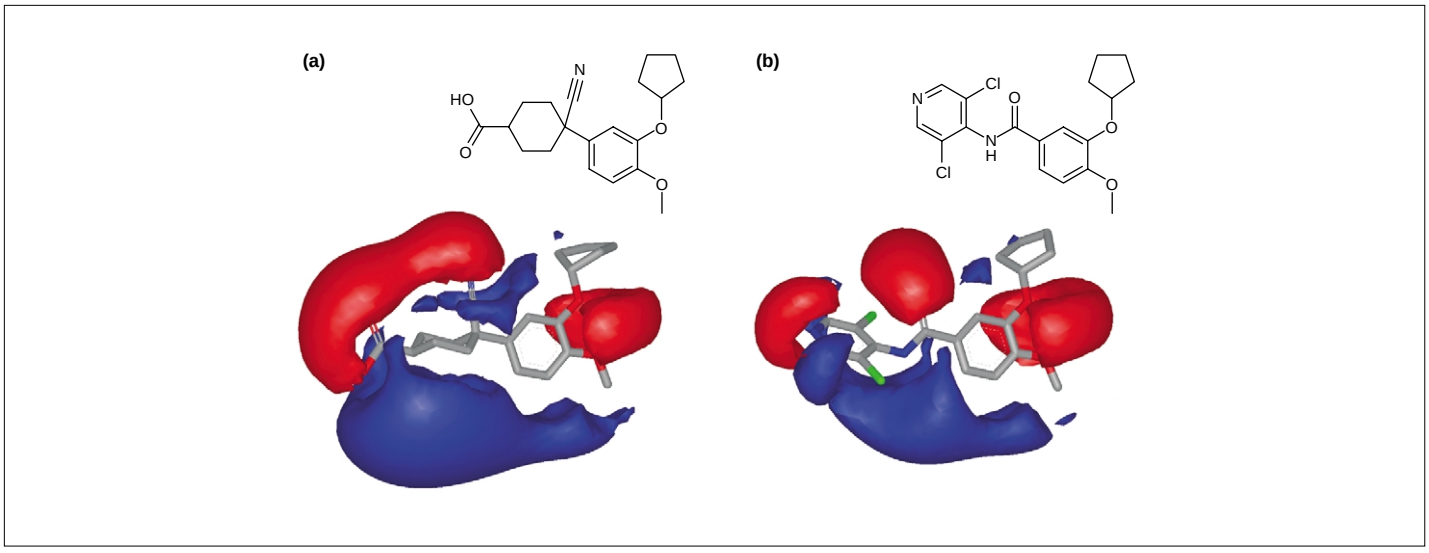


FIGURE 4

Database search based on shape and electrostatic similarity can identify structures that have similar chemical properties (electrostatics) to a template molecule but are structurally dissimilar. The two inhibitors of Phosphodiesterase 4B, **(a)** cilomilast and **(b)** piclamilast, are structurally different but have high shape and electrostatic similarity.

and no conditions on types of molecules or types of chemistry. 3D methods obviously have their particular problems, most notably the multiple conformer issue. However, shape-comparison programs alone work surprisingly well. For example, Rush *et al.* [37] of Wyeth recently used the shape comparison program ROCS (OpenEye Scientific Software, www.eyesopen.com) to identify a set of novel inhibitors of the ZipA–FtsZ protein–protein interaction. The obvious advancement of this approach is to use shape in combination with electrostatics. The clomplementary program to ROCS for electrostatics is the recently developed program EON (also by OpenEye Scientific Software). EON calculates the electrostatic field around aligned molecules using Poisson–Boltzmann theory (Figure 4). Nicholls and co-workers found that molecular shape and electrostatics, in combination with 2D structural fingerprints are important variables in discriminating between active and inactive compounds [38]. A related – although not shape-based – method is implemented in Cresset FieldScreen™ software (www.cresset-bmd.com), which is based on molecular field points around a molecule. Low *et al.* [39] applied this tool to derive novel selective cholecystokinin-2 (CCK_2) antagonists.

Data mining: use knowledge to find actives

The techniques described above are very efficient in identifying compounds that are similar to a given hit or representative of certain hit series. Data-mining approaches [40] can take into account global knowledge, such as features present not only in an interesting hit series but in any of the compounds tested up to a point. In these tools, available actives and inactives are used as training sets to build models capable of classifying new compounds by labelling them as active or inactive, or by providing a score related to the probability of these compounds being active.

Data mining has become widely known within drug discovery by a fairly large number of approaches to the prediction of drug likeness, for example, by using artificial neural networks (ANNs) [41–44], recursive partitioning [45] or support vector machines (SVMs) [46]. Thousands of representatives are accessible for model building [e.g. the MDL Drug Data Report (MDDR)] for predicting drug likeness. However, during hit-to-lead only a few hundred actives are available. Also, a much higher number of inactives than potential actives is expected in any set that is to be classified, and models that reflect this skewed class distribution need to be developed. These models should be more accurate in classifying inactives to reduce the total number of compounds that are predicted to be active and minimize the false-positive rate [47].

There have been successful reports of data mining being used for selecting focused screening sets. Neural networks have been trained to predict target-class likeness for, for example, G-protein-coupled receptors (GPCRs) or kinases [14,48]. Similarly, Saeh *et al.* [47] of AstraZeneca used SVMs to predict activity against specific targets. Here, the molecules are represented by 3D pharmacophoric fingerprints. Because of this, the model was able to identify chemical classes not represented in the training set, which proves that classification can even be used for lead hopping. Recursive partitioning [49] has been used to select subsets during sequential screening for GPCRs [13]. Another interesting approach is described by Warmuth *et al.* [50], who use SVM and active learning during sequential screening. During active learning, the model is refined after each iterative screening step, and the compounds to be screened are chosen by a particular strategy. Warmuth *et al.* propose a way of switching strategies for biasing the outcome to exploration or exploitation of the chemical space.

Just as in the field of fingerprint similarity searching and clustering, several comparisons between data-mining methods have been carried out. It is not possible to reach a consensus from these studies because performance clearly depends on the composition and size of training sets. When the number of compounds in the training sets is not much larger than the number of descriptors, there is the risk of overfitting [51]. In such cases, feature selection should be used to reduce the number of descriptors [52]. Overfitting needs to be carefully monitored, especially for ANNs, because these minimize only the error in the training data during model building. SVMs are based on the concept of structural risk minimization, in other words, they take into account the approximation error during

model building, which makes them less prone to overfitting [53]. SVMs have been shown to outperform ANNs in a range of test sets [46,54,55]. However, the observed differences are often minor.

In the past few years, several data-mining techniques have been improved by making use of ensembles of classifiers that are then used to derive a consensus model. This was first introduced for recursive partitioning [10,56,57] and has more recently been applied to SVM classification [58]. There is a significant performance gain when the ensemble methods are applied to known datasets and compared to the performance of the original non-ensemble classification. An interesting example along this line is the prediction of isoform selectivity of UDP-glucuronosyltransferase by Sorich *et al.* [54]. Following up on earlier work using 2D descriptors and SVM, the authors add quantum-chemical descriptors and show that the results improve significantly when building separate SVM models and using a consensus classification rather than combining both descriptor sets in a single model [59].

Prioritization: ranking hit series

We have described the computational tools used to explore the environment of compounds to enrich hit series so that they show initial SAR. Before lead-identification chemistry starts, the series need to be profiled further to explore their potential as lead series.

Ligand efficiency [60,61] is currently one of the most frequently used buzzwords in hit-to-lead. The idea behind it is to look at potency in combination with size, so that small, less potent molecules are ranked equally with large, more potent compounds. This makes it a very useful tool to compare compounds across different hit series or to pinpoint small interesting clusters during hit identification, which can be easily overlooked when focusing on potency alone. However, it is important that ligand efficiency should always be seen in the context of a set of compounds, because it is not a global measure for the overall quality of a hit. High ligand efficiency for a small weakly active molecule does not necessarily imply that it has the potential to be turned into a high affinity compound[†]. Most desirable are hit series where potency can be increased while maintaining molecular weight. Increasing molecular weight is obviously not the only way to increase potency but Oprea *et al.* [62] argue that drugs are on average substantially larger than leads.

The rapid delivery of a compound that can be used as a pharmacological tool has the highest priority for non-validated targets. In this case, potency and physicochemical properties are clearly more important than ligand efficiency. Lead series should provide orally available drug molecules by the end of most drug discovery projects. Oral bioavailability depends on several physicochemical properties, and Lipinski's rules are widely accepted as a general filter. It has recently been shown by Martin [63] of Abbott that these rules need to be further refined, especially for charged compounds. Even when the final goal is an orally available drug, there is no generic profile for an optimal hit series because the optimal properties depend on the target class or therapeutic indication. For example, drug molecules that target a nuclear receptor need to meet a different profile than those that block a receptor at a cell surface.

Important for ranking hit series are their absorption, distribution, metabolism and elimination (ADME) properties and toxicity (Tox) profiles. ADME–Tox properties, such as solubility [64], metabolic stability [65], permeability [66], plasma-protein binding [67], toxicity [68] and cytochrome P450 inhibition [69], can be measured experimentally, or be based on predictive models [70,71]. Both alternatives should complement each other: computational models always need experimental data for refinement, and if a reliable[‡] model is available, there is no need to perform an experiment. In general, the ADME–Tox profiles should be evaluated together with potency and SAR information and enable project teams to make informed decisions, to see the potential of a series, and to be alerted to potential show stoppers [72].

Hit-to-lead: making the right choices

In a very entertaining article, DeWitte [73] compared the difficulties of sequential selections during drug discovery with the sports world, where occasional victories should not outperform consistent success. As an example, he claims that the best cyclists would never have been discovered if during the Tour de France only the top cyclists from the flat races were allowed to race across mountainous terrain. Because we will never be able to carry out *in vivo* experiments with all our hits, the main challenge during hit-to-lead is to make the right decisions at the various stages and, while doing so, not to lose sight of the larger context. Instead of being impressed by the high potency or high ligand efficiency of single hits, one should always try to find the right balance between potency, SAR and pharmacological profile. All available information must be considered to identify the potential of a series being successful in future experiments.

The overall lead generation process starts with a large number of primary actives, but many of them can turn out to be false positives. False negatives can also occur, most likely because of sample decomposition or precipitation [74]. In the early stages of hit identification, decisions need to be based on statistics. No data can be judged alone and it is important that the subsequent hit-to-lead process starts with quality-controlled hits. These hits are organized in clusters and it can be worthwhile to screen neighbours to original hits in a second round to enrich smaller clusters or to follow up singletons. Several techniques for this task have been described above, and we must stress once more that one should combine as many different computational methods for searching as possible. In addition, it is important to be more open in the beginning – to explore more clusters of hits – before focusing on a few series. Although we described most techniques in the context of HTS follow-up, these are all applicable to selecting subsets for focused screening.

When comparing hit series, it is important to grasp the overall potential of a series, and therefore useful to also carefully look at the inactives when analysing the SAR. Regarding the ADME–Tox profile of a series, it is not expected that the perfect compound with oral bioavailability will be found at this stage. There might be shortcomings in potent compounds that are not present in other, less potent members of the series. The hope is that whatever issue there is, this is not related to the chemical class of the series (or even the target) but only observed in a few members. By working closely with medicinal chemists in several projects, we realize that there is a significant 'human factor' in choosing which series to

[†]A water molecule would be the optimal starting point according to this: it always binds and has low MW.

[‡]'All models are wrong – but some are useful.' (Statistician George P. E. Box).

take further. Synthetic accessibility is a very important criterion. The easier it is to quickly explore a series by making a library of desired analogues in a few steps, the higher it is ranked by the chemist. The chemist's background (i.e. favourite reactions) is obviously important here. This is pragmatic and easy to justify. Another experience, which is perhaps more difficult to defend, is that medicinal chemists tend to find the biggest clusters the most interesting to follow up, and most programs also rank larger clusters higher than smaller ones. Although it is certainly reassuring to see that a large number of compounds in a series show activity, it is more important that initial SAR can be derived from the available compounds.

Conclusion

During hit identification and the subsequent hit-to-lead process, a large number of initial actives are followed up and prioritized. Cheminformatics and computational chemistry tools are used at all stages to provide an overview of the hits, organize them into series and search for analogues to enrich these series. Different computational methods are available, and there is no single generic workflow, but it is the combination of several techniques that provides the best results. All methods are well established, especially large-scale property prediction, which now provides better support during decision-making than ever before. However, the computer can just prepare the data, in the end it is up to the project team to make use of all available information to ensure that the most promising series are progressed into lead optimization.

Acknowledgements

The authors thank Tim Perkins and Kay Brickmann for discussions and for their useful feedback on the manuscript.

References

1 Davis, A.M. *et al.* (2005) Components of successful lead generation. *Curr. Top. Med. Chem.* 5, 421–439
2 Bleicher, K.H. *et al.* (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378
3 Congreve, M. *et al.* (2005) Keynote review: structural biology and drug discovery. *Drug Discov. Today* 10, 895–907
4 Lloyd, D.G. *et al.* (2004) Scaffold hopping in *de novo* design. ligand generation in the absence of receptor information. *J. Med. Chem.* 47, 493–496
5 Dror, O. *et al.* (2004) Predicting molecular interactions *in silico*: I. a guide to pharmacophore identification and its applications to drug design. *Curr. Med. Chem.* 11, 71–90
6 Posner, B.A. (2005) High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug Discov. Devel.* 8, 487–494
7 Golebiowski, A. *et al.* (2001) Lead compounds discovered from libraries. *Curr. Opin. Chem. Biol.* 5, 273–284
8 Golebiowski, A. *et al.* (2003) Lead compounds discovered from libraries: Part 2. *Curr. Opin. Chem. Biol.* 7, 308–325
9 Fox, S. *et al.* (2004) High-throughput screening: searching for higher productivity. *J. Biomol. Screen.* 9, 354–358
10 van Rhee, A.M. (2003) Use of recursion forests in the sequential screening process: consensus selection by multiple recursion trees. *J. Chem. Inf. Comput. Sci.* 43, 941–948
11 Karnachi, P.S. and Brown, F.K. (2004) Practical approaches to efficient screening: information-rich screening protocol. *J. Biomol. Screen.* 9, 678–686
12 Shanmugasundaram, V. *et al.* (2005) Hit-directed nearest-neighbor searching. *J. Med. Chem.* 48, 240–248
13 Jones-Hertzog, D.K. *et al.* (1999) Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J. Pharmacol. Toxicol. Methods* 42, 207–215
14 Ford, M.G. *et al.* (2004) Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks. *J. Mol. Graph. Model.* 22, 467–472
15 Clark, D.E. *et al.* (2004) A virtual screening approach to finding novel and potent antagonists at the melanin-concentrating hormone 1 receptor. *J. Med. Chem.* 47, 3962–3971
16 Gribbon, P. and Andreas, S. (2005) High-throughput drug discovery: what can we expect from HTS? *Drug Discov. Today* 10, 17–22
17 Fischer, H.P. and Heyse, S. (2005) From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr. Opin. Drug Discov. Devel.* 8, 334–346
18 Roche, O. *et al.* (2002) Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* 45, 137–142
19 McGovern, S.L. *et al.* (2003) A specific mechanism of nonspecific inhibition. *J. Med. Chem.* 46, 4265–4272
20 Engels, M.F.M. *et al.* (2002) Outlier mining in high throughput screening experiments. *J. Biomol. Screen.* 7, 341–351
21 Schreyer, S.K. *et al.* (2004) Data shaving: a focused screening approach. *J. Chem. Inf. Comput. Sci.* 44, 470–479
22 Glick, M. *et al.* (2004) Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screen.* 9, 32–36
23 Diller, D.J. and Hobbs, D.W. (2004) Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* 47, 6373–6383
24 Stahura, F.L. and Bajorath, J. (2003) Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* 10, 707–715
25 Wilkens, S.J. *et al.* (2005) HierS: hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* 48, 3182–3193
26 Stahl, M. *et al.* (2005) A robust clustering method for chemical structures. *J. Med. Chem.* 48, 4358–4366
27 Stahl, M. and Mauser, H. (2005) Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Comput. Sci.* 45, 542–548
28 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903–911
29 Ginn, C.M.R. *et al.* (2000) Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Des.* 20, 1–16
30 Salim, N. *et al.* (2003) Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* 43, 435–442
31 Martin, Y.C. *et al.* (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358
32 Fowler, D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 38, 379–386
33 Holliday, J.D. *et al.* (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* 43, 819–828
34 Hert, J. *et al.* (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* 44, 1177–1185
35 Whittle, M. *et al.* (2004) Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* 44, 1840–1848
36 Grant, J.A. *et al.* (1996) A fast method of molecular shape comparison: a simple application of a gaussian description of molecular shape. *J. Comput. Chem.* 17, 1653–1666
37 Rush, T.S. 3rd *et al.* (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* 48, 1489–1495
38 Nicholls, A. *et al.* (2004) Variable selection and model validation of 2D and 3D molecular descriptors. *J. Comput. Aided Mol. Des.* 18, 451–474
39 Low, C.M.R. *et al.* (2005) Scaffold hopping with molecular field points: identification of a cholecystokinin-2 (Cck2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based Cck2 antagonists. *J. Med. Chem.* 48, 6790–6802
40 Weaver, D.C. (2004) Applying data mining techniques to library design, lead generation and lead optimization. *Curr. Opin. Chem. Biol.* 8, 264–270
41 Ajay, A. *et al.* (1998) Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* 41, 3314–3324
42 Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 41, 3325–3329
43 Frimurer, T.M. *et al.* (2000) Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* 40, 1315–1324
44 Murcia-Soler, M. *et al.* (2003) Drugs and nondrugs: an effective discrimination with topological methods and artificial neural networks. *J. Chem. Inf. Comput. Sci.* 43, 1688–1702
45 Wagener, M. and van Geerestein, V.J. (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* 40, 280–292

46 Zernov, V.V. *et al.* (2003) Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* 43, 2048–2056

47 Saeh, J.C. *et al.* (2005) Lead hopping using SVM and 3D pharmacophore fingerprints. *J. Chem. Inf. Model.* 45, 1122–1133

48 Manallack, D.T. *et al.* (2002) Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* 42, 1256–1262

49 Rusinko, A. *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026

50 Warmuth, M.K. *et al.* (2003) Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* 43, 667–673

51 Hawkins, D.M. (2004) The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12

52 Walters, W.P. and Goldman, B.B. (2005) Feature selection in quantitative structure-activity relationships. *Curr. Opin. Drug Discov. Devel.* 8, 329–333

53 Kecman, V. (2001) *Learning and Soft Computing*, The MIT Press

54 Sorich, M.J. *et al.* (2003) Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* 43, 2019–2024

55 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

56 Tong, W. *et al.* (2003) Decision forest: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* 43, 525–531

57 Svetnik, V. *et al.* (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958

58 Merkwirth, C. *et al.* (2004) Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* 44, 1971–1978

59 Sorich, M.J. *et al.* (2004) Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J. Med. Chem.* 47, 5311–5317

60 Hopkins, A.L. *et al.* (2004) Ligand Efficiency: A useful metric for lead selection. *Drug Discov. Today* 9, 430–431

61 Abad-Zapatero, C. and Metz, J.T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* 10, 464–469

62 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315

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

64 Delaney, J.S. (2005) Predicting aqueous solubility from structure. *Drug Discov. Today* 10, 289–295

65 Nassar, A-E.F. *et al.* (2004) Improving the decision-making process in the structural modification of drug candidates: enhancing metabolic stability. *Drug Discov. Today* 9, 1020–1028

66 Malkia, A. *et al.* (2004) Drug permeation in biomembranes: *in vitro* and *in silico* prediction and influence of physicochemical properties. *Eur. J. Pharm. Sci.* 23, 13–47

67 Kratochwil, N.A. *et al.* (2004) Predicting plasma protein binding of drugs – revisited. *Curr. Opin. Drug Discov. Devel.* 7, 507–512

68 Nassar, A-E.F. *et al.* (2004) Improving the decision-making process in structural modification of drug candidates: reducing toxicity. *Drug Discov. Today* 9, 1055–1064

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

70 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modeling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204

71 Beresford, A.P. *et al.* (2004) In silico prediction of ADME properties: are we making progress? *Curr. Opin. Drug Discov. Devel.* 7, 36–42

72 Kerns, E.H. and Di, L. (2003) Pharmaceutical profiling in drug discovery. *Drug Discov. Today* 8, 323

73 DeWitte, R.S. (2002) On experimental design in drug discovery. *Curr. Drug Discovery* February, 19-22

74 Cheng, X. *et al.* (2003) Studies on repository compound stability in DMSO under various conditions. *J. Biomol. Screen.* 8, 292–304