The impact of informatics and computational chemistry on synthesis and screening

Charles J. Manly, Shirley Louise-May and Jack D. Hammer

High-throughput synthesis and screening technologies have enhanced the impact of computational chemistry on the drug discovery process. From the design of targeted, drug-like libraries to 'virtual' optimization of potency, selectivity and ADME/Tox properties, computational chemists are able to efficiently manage costly resources and dramatically shorten drug discovery cycle times. This review will describe some of the successful strategies and applications of state-of-the-art algorithms to enhance drug discovery, as well as key points in the drug discovery process where computational methods can have, and have had, greatest impact.

***Charles J. Manly**
Shirley Louise-May
and Jack D. Hammer
Neurogen Corporation
35 Northeast Industrial Rd
Branford, CT 06405, USA
*tel: +1 203 488 8201
fax: +1 203 481 5290
e-mail:cmanly@nrgn.com

▼ Informatics and computational chemistry have the potential to play a tremendously important and diverse role in drug discovery, and virtual screening is an important focus of these disciplines. Virtual screening[1] is the term broadly used to mean the computational analysis of databases of chemical compounds to identify possible drug candidates for a specific pharmaceutical target. Many virtual screening efforts focus exclusively on the use of three-dimensional (3D) macromolecule binding-site information[2,3] but a large number of membrane-bound targets lack macromolecular binding-site data. In fact, membrane bound receptors, such as ion channels and G-protein-coupled receptors (GPCRs), represent ~50% of current drug targets and remain mostly beyond the aid of these explicit 3D technologies[4]. Hence, virtual screening has come to include alternative approaches that do not explicitly rely on macromolecular binding-site information[5]. The focus of this review is a description of these approaches and the impact they have had, with an emphasis on small molecule strategies. The literature of structure-based drug design and macromolecular structure-based approaches has been reviewed recently[1,2,6].

Process issues in modern drug discovery

Computational chemistry approaches that optimize the prioritization and focus of high-throughput methods have a powerful impact when deployed knowledgably[7]. Computational chemistry and informatics can be used to integrate workflow and dataflow for optimum effectiveness in achieving project goals. This integration makes possible fast iterative virtual screening, to effectively prioritize targeted synthesis and screening efforts.

There are several other key opportunities for computational chemistry intervention in modern drug discovery. One of the most important of these is efficient target validation. With the coming wave of information from the Human Genome Project, a large number of therapeutic targets for pharmaceutical and biotechnology companies is anticipated. However, the validation of these as viable therapeutic targets will take considerable time and effort, and there will probably be an appreciable lag between identifying new receptor targets and obtaining structural information for these new targets. The recent history of orphan receptors and the time and effort expended to validate them as therapeutic targets indicates the magnitude of the validation effort that lies ahead[4]. This leaves the significant problem of how to identify useful leads efficiently[3]. Computational chemistry approaches that do not rely on receptor structure information are, therefore, likely to become a more central strategy of computational chemistry groups. As drug discovery research moves from an era characterized by

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(01)01990-0

fewer targets with a higher degree of validation, to a future of many more targets with a much lower degree of validation, the ability to quickly and efficiently pursue multiple drug targets – in parallel with the efforts to validate them – will be a strong asset to drug discovery efforts[5].

Another key point of intervention is lead identification. The goals of computational chemistry modeling in drug discovery have usually focused on the ability to design active molecules when given an adequate model. Growing numbers of companies have found that solving this very difficult problem becomes somewhat easier by re-phrasing the problem as one of predicting which molecules from a large set of compounds – a virtual library – will possess highest activity. Although this might seem a simple semantic difference, the simplifications brought by re-phrasing the problem in this way are truly significant.

Because computational chemistry can look beyond the activity of individual molecules and analyze activities across populations of molecules, it can mine new information from the analysis of HTS data on inactive and weakly active compounds that were previously ignored, as well as analysis of data that have more noise than is desirable. This is feasible because of new model construction technologies, such as machine learning methods[8]. Data quality in HTS efforts has also improved such that all data, including those associated with weakly active and inactive compounds, are being used in subsequent analysis[9]. The role of HTS has thus progressed from simple scanning for hits to generating data for directing further screening and synthesis efforts, and for lead identification as well as optimization. Perhaps the most important difference, however, is simply the conscious decision to focus computational chemistry efforts in this way. At Neurogen (Branford, CT, USA) and in other drug discovery efforts, this focus is developing SARs earlier in projects, often before any synthetic exploration on the project has begun.

High-throughput screening and high-throughput data

Historically, HTS efforts in pharmaceutical discovery were used as a filter to identify the few potentially promising hits for further analysis in 'campaign' screening of a corporation's entire synthetic archive. The bulk of the real data produced using the HTS effort was not used further, and the data produced on weakly active or inactive compounds were ignored. In fact, even the hits from combinatorial chemistry libraries were sometimes ignored because – in the eyes of the medicinal chemist – they were seen as lacking drug-like qualities. Both HTS and combinatorial chemistry efforts often existed separately and were isolated from what were considered the main efforts in drug discovery: generating leads from the hits and then optimizing these toward clinical candidacy. By contrast, HTS, combinatorial chemistry and computational chemistry are now being used to direct individual screening runs to optimize the information content and activity of the results. These efforts encourage 'targeted sequential screening' as the alternative to campaign screening, and can be efficient at mining the active compounds from the library early in the project[10,11].

The desire to use HTS-generated data as more than simply a filter, together with the increasing throughput possible in modern assay technologies, requires fast and effective approaches for the analysis of these data. This has provided new opportunities for informatics and computational chemistry to have impact and also has the effect of transforming HTS from a tool for simply identifying hits, to a process for providing a continuum of validated information about compounds in a project. Such information now includes cell-based functional efficacy, selectivity, side-effect profiling, certain *in vitro* ADME/Tox properties, pharmacokinetic data and data relating to drug-likeness.

Combinatorial chemistry and parallel synthesis (high-speed synthesis)

Combinatorial chemistry has been a part of drug discovery efforts for almost a decade, but its impact is less than was originally envisioned. Combinatorial synthesis and high-speed synthesis (HSS) first focused on the quantity and speed of synthesis and, hence, stressed what could be made rather than what should be made. Historically, this was because of the limited synthetic protocols available, the extensive chemistry development necessary for a new protocol, and the belief that simply generating large libraries would successfully address drug discovery problems.

Although much effort still focuses on what can be synthesized, targeted and directed library development is increasingly being used[1,12,13]. As efforts are directed toward generating large numbers of reactive fragments and developing new synthetic protocols, more significant chemical space becomes accessible by modern high-speed solution and solid-phase synthetic technologies. In addition, medicinal chemistry concepts are being increasingly considered in library design, and computational chemistry can provide the tools to accelerate the integration of these concepts through the development and application of models of drug-like properties and SARs. With the integration of computational chemistry and synthesis via the virtual library and virtual screening, computational chemistry has the opportunity for greater impact than ever before.

Compound attrition in the drug discovery pipeline

Pharmaceutical companies are increasingly concerned about managing their drug discovery efforts appropriately.

Current estimates suggest that ten compounds must enter the development pipeline for every new drug application (NDA), and managing the attrition curve has become of paramount importance[14]. Considering the failure rates between each step of the process, it can clearly be seen that early intervention has great potential to change the overall numbers dramatically and beneficially.

Poor biopharmaceutical properties in candidate compounds contribute strongly to both failures and development slowdowns[14], as can be seen from the disappointing performance of compounds derived from simple HTS and combinatorial efforts[6]. These failures can, in part, be attributed to focusing these HTS and HSS efforts on identifying and optimizing leads while at the same time ignoring less active leads that already possess reasonable starting points for ADME/Tox properties. It is probable that the difficulty of optimizing a lead with poor ADME/Tox and drug-like properties has been underestimated in the decision processes of both lead candidate selection and development[14,15].

For this reason, early assessment of ADME/Tox properties is receiving considerable attention and resources[16,17]. Computational chemistry approaches are playing key roles in the assessment of these liabilities, and will significantly impact the success rate at each step of the development process. These approaches will be discussed in more detail later[13,18–22].

Integration of drug discovery through informatics

The drug discovery cycle of screening, analysis, design, synthesis and screening has been the cornerstone of pharmaceutical discovery for years. Each cycle produces improved chemistry, guided by new biological information and SARs. Informatics is accelerating this cycle through the tight integration of individual efforts and disciplines in drug discovery, especially the high-throughput disciplines. In many environments, the cycle time has decreased to days or weeks, instead of months; this is the result of integrating and streamlining the dataflow and workflow between disciplines.

Informatics efforts at many companies efficiently collate, analyze and convey SAR knowledge to scientists, enabling them to direct their projects more effectively. The trend toward even greater impact is promised, as informatics sophistication increases.

High-throughput computational modeling concepts

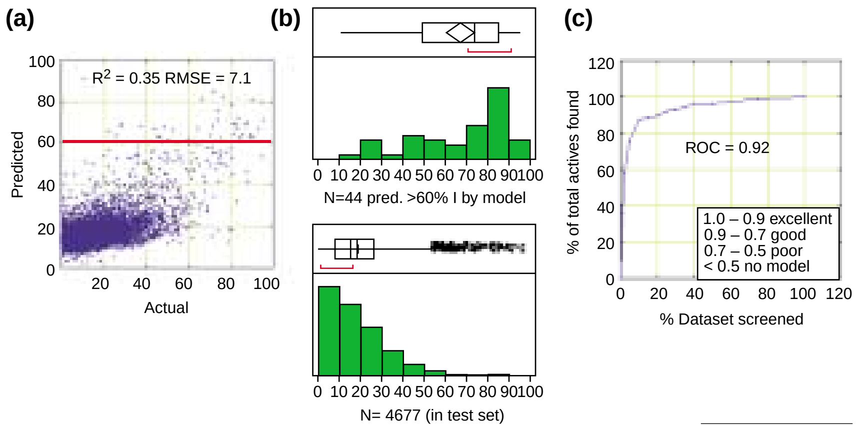
Current computational algorithms are moving beyond the design of combinatorial libraries merely for diversity, and virtual screening of corporate archives solely to identify hits, and are moving toward 'virtual' optimization of hits in terms of selectivity, and drug-like and ADME/Tox properties[23–29]. Computational filters, encoding 'rule of five' principles[30] and lists of toxophores[31], can be applied to address the drug-like properties and toxicity liabilities of potential libraries before synthesis. Addressing these characteristics early in the discovery process can improve hit quality, and provide the opportunity for their optimization in parallel with activity optimization. Most prominently, virtual screening of compound libraries using computational filters and models enables focused consideration of smaller, biologically relevant and activity enriched portions of the available chemical space.

Model evaluation

The success of high-throughput computational models in virtual screening can be affected significantly by the modeling protocol, that is, the complementarity between the target property being modeled, the modeling method, and the choice of molecular representation (descriptor sets). There are presently dozens of modeling technologies[1,32] that can be paired with a multitude of descriptor sets[33]. This has necessitated the development of tools that assess model predictiveness. Fortunately, the sheer volume of HTS data available as input for computational models allows for routine evaluation of these models.

Figure 1 illustrates several analyses of a test set of 20% of an HTS binding affinity screen used to evaluate the predictions from a neural network model built on the remaining 80% of data (data generated at Neurogen Corporation, 2001). The correlation plot in Figure 1a, of predicted versus actual activity values, is a simple way to graphically depict model behavior. However, the model's use is more clearly demonstrated by comparing the population of the test set as a whole (Figure 1b, bottom) with the subset of compounds with predicted activity of >60% by the neural net model (Figure 1b, top). The activity distribution of the subset has shifted significantly towards higher activity by application of this model. The enrichment afforded by a model can be quantified and normalized by determining the area under its receiver operating characteristic (ROC) curve[34], a curve that describes the percentage of actives found as a function of the percentage of a library screened (Figure 1c).

The inability to sample a significant portion of biologically relevant chemical space adequately, even with high-throughput efforts in synthesis and screening, limits the predictiveness of computational models. Characterization of model coverage to identify sparsely sampled or unsampled regions of chemical space provides a valuable estimate of the confidence of model predictions. For example, some form of k-nearest neighbor analysis[35] can be used to determine



Drug Discovery Today

Figure 1. Computational model evaluation. Panel **(a)** shows a correlation plot (real versus predicted) for a test set with a prediction threshold at 60% (red line). Panel **(b)** shows the probability shift and activity distribution plots for compounds predicted >60% (top), and all compounds in test set (bottom). Panel **(c)** shows the receiver operating characteristic (ROC) curve (or 'lift curve') with integration area under the curve.

where a predicted compound resides in descriptor space, relative to the compounds that comprise the model. If the Euclidean distances (straight-line distance in descriptor space) between the predicted compound and its nearest neighbors in the model are relatively large, then that compound's prediction would be treated as an extrapolation. The development of these types of proximity metrics[36] are needed to more effectively couple computational modeling with virtual screening of extremely large virtual libraries.

Modeling techniques

Currently, high-throughput computational models can be designed to assess similarity or diversity[37,38]; to predict biological activity[39], selectivity[40] or toxicity[31]; or to discriminate size, shape, pharmacophores and related 3D properties[41]. These methods can be used with a growing number of descriptor types and can be applied to an increasing range of target properties[42]. There are several 3D quantitative SAR (QSAR) and pharmacophore methodologies and some of these have been extended to higher dimensions. Although they are very powerful techniques[43,44], they rely on the generation of a reasonable molecular conformation(s), which limits their application in virtual screening of large numbers of compounds ($>10^8$), or dynamic or un-enumerated virtual libraries, and thus will not be covered here.

Optimization of the modeling protocol and descriptor feature selection have been the subject of recent developments in high-throughput modeling technology. Multivariate adaptive regression splines (MARS), a statistical method of function approximation developed in 1991 (Ref. 45), has recently been applied to data mining of drug discovery targets[46]. This method enables the detection of subtle descriptor interactions, as well as optimization of a model to adjust for these interactions[47]. Bayesian regularized artificial neural networks, used in conjunction with automatic relevance determination (BRANN-ARD; Ref. 48), successfully addresses several limitations in the application of artificial neural networks to drug discovery problems (i.e. the need for validation sets to check against model overtraining, the inability to interpret model results in terms of the input descriptors, objective descriptor feature selection and confidence assessments of the model predictions[48]).

Descriptors

The multitude of descriptor sets available for use in high-throughput computational modeling can be roughly classified by the dimensionality of the molecular representation from which they are derived. Because the simplest descriptor types have been applied successfully to filtering large libraries, whereas the most complex have been applied primarily to small subsets of libraries, their application can be further characterized by a pyramid, as described in the next section.

Descriptor dimensionality pyramid

Bottom level: 1D – physicochemical and bulk properties. 1D descriptors, such as molecular weight and logP, are rapidly calculable and reflect general compound properties, such as size and lipophilicity. They are relatively few in number and are generally fragment additive. They are often applied as coarse-grained filters to improve the drug-likeness of libraries in the form of Lipinski's 'rule of five' criteria for improved oral bioavailability[30].

Mid level: 2D – structural fragment and topological properties. 2D descriptor sets, because of their dimensionality $O(10^2–10^4)$, are often referred to as molecular fingerprints. Substructure searches of predefined fragment lists[49] can be performed efficiently on SMILES (Simplified Molecular Input Line Entry System; Daylight Toolkit Software, Daylight Chemical Informations Systems, Mission Viejo, CA, USA) representations, and topological indices are readily calculable from molecular connectivity tables (e.g. Molconn-X Version 2.0, Hall Associates Consulting, Cambridge, MA, USA). Although substructure fragments are inherently adaptable to fragment methods, topological descriptors can also be adapted to fragment methods by allowing overlap between fragments that aid in the continuity of the topological concepts. Data mining techniques often employ topological or substructure descriptor fingerprints in modeling of binding affinity[33,50], in evaluation of toxicity liabilities[31], and in analyses of diversity for design of combinatorial libraries or evaluation of external libraries[21,51].

Upper level: 3D – surface, shape and volume properties, fields and spatial pharmacophores. 3D descriptors have been developed to capture explicitly the discriminating and specific nature of ligand–receptor interactions. However, issues of conformational generation, sampling and alignment have impeded their general use in high-throughput computational methods[52,53]. Current developments of 3D descriptors circumvent these issues by several strategies. Fragment-based and connectivity-based calculations of surface area related descriptors have been developed that compare well with the explicit 3D calculation[54,55], and extensive use has been made of polar surface area (PSA) and related descriptors to model transport properties[56,57]. Topomers are a set of predefined shape and/or volume fragment descriptors analogous to 2D substructure fragments, that have been successfully applied to similarity searching of a large virtual library for angiotensin II antagonists[37,38].

The presence or absence of spatial pharmacophores, or the through-space relationship of functional groups relevant to binding affinity, can be used to construct a conformation-sensitive fingerprint. Mason and coworkers, in their construction of four-point pharmacophore fingerprints, constrain the fourth 'point' to be a feature unique to the privileged structure fragment that is part of the pharmacophore definition[58]. (Privileged structures are molecular substructures that are common to high-affinity ligands of distinct receptors[59].)

Unexpectedly, 2D substructure fragment fingerprints have performed consistently as well as, or better than, the more costly 3D descriptors[42,50], and seem to implicitly encode at least some degree of physicochemical properties, atom types, topology, shape and surface properties[42]. However, some of this behavior could result from the predominance of structurally congeneric series (i.e. related in a 2D sense) in the training sets and test sets published in the medicinal chemistry literature.

Virtual libraries and virtual screening

A variety of high-throughput computational methods has been developed and applied to library design, library filtering and virtual screening. The goal of library design is to construct a population that is as large and relevant as possible. Library filtering rapidly removes from further consideration any molecules in the library that are outside a minimum set of practical conditions, for example drug-likeness, complexity, synthetic feasibility and cost. Virtual screening of the remaining library using a computational model ranks, orders and thus focuses synthesis and screening efforts on the segments of the library that are most relevant to the project goals.

Library design

Retrospective analyses of diversity-designed combinatorial libraries[30] reveal product distributions that were higher in molecular weight and in the number of rotatable bonds[11], and that were biased toward significantly higher or lower lipophilicity values than known, orally active drugs[27]. As a consequence, hits from HTS screens of combinatorial libraries often presented nontrivial solubility and permeability hurdles. Although diversity remains an integral part of library design, it is now balanced by strategies more relevant to the drug discovery process[6], primarily the design of focused libraries[24], 'drug-like' libraries[21,28], 'CNS-like' libraries[60], the exclusion of compounds with toxophores[31], and the inclusion of practical considerations[25,27]. An extensive analysis of existing drug profiles has helped to guide these design considerations[28] and the realization that leads differ from drugs in property space has also recently been highlighted[25,61].

A new theme emerging in the design of combinatorial libraries uses the concept of the privileged structures (discussed previously) for a given class of receptors, such as GPCRs (Ref. 58). The evidence of privileged structures suggests that, although theoretical chemical space is nearly infinite, biological activity in that space might be relatively conservative, occurring in clusters or neighborhoods around substructure elements[59]. PLUMS (Ref. 24) is a rational monomer selection method that incorporates the ability to optimize 2D and 3D property criteria in the design of focused libraries that can also accommodate the privileged structure concept.

The incorporation of medicinal chemistry concepts and, more importantly, the involvement of medicinal chemists in combinatorial library design[62], is required to advance HSS impact beyond satisfaction of diversity and throughput expectations. Design of combinatorial libraries with respect to pharmaceutically relevant criteria could result in HSS hits that are much closer to clinical candidacy.

Project modeling strategies

As a drug discovery project progresses from lead generation to candidate selection, the computational tools needed to make an impact change. At the beginning of a project, structures from the literature can be used as similarity searching probes[5]. By contrast, if nothing is known about the SAR initially, screening a structurally diverse subset of an archive can provide a starting point. If weak to moderate activity is found, computational tools such as similarity searching[42] and/or self-organizing maps[8,63] can be used to mine the library around these hits. The compounds that result from this approach can then be tested during subsequent rounds of screening. Usually, after just a few rounds of the data mining and screening cycle, enough active compounds will have been found to derive a model. Even weakly predictive models evolve over time and can become quite powerful in terms of hit rate enhancement.

Portfolio deployment

There is tremendous value in deploying multiple modeling protocols. As a discovery project evolves, the optimal combination of modeling technology and descriptor type can easily change. Using multiple modeling protocols from the beginning can greatly increase the chances of having a predictive model at any given time. Furthermore, although two different modeling protocols might find many of the same active compounds and series within a given library, each modeling protocol could also identify active series not found by the other. In other words, there can be a substantial degree of complementarity between modeling approaches.

Targeted sequential screening

Figure 2 depicts the sequential process of virtual screening of a library. From an initial pool of data, a model is constructed. Then, the top-predicted compounds from the first virtual-screening cycle are tested in the assay and these new data are added to the original data set. The augmented data set is then used to construct a new model, which is used to virtually screen the remainder of the library where, again, the top-predicted compounds are selected for screening. If the compound library is reasonably well-suited to the therapeutic target, several cycles of virtual and actual screening[64,65] can often lead to the identification of significant numbers of lead compounds in the submicromolar and nanomolar range (data from Neurogen Corporation). This approach is particularly powerful when computational and medicinal chemists coordinate their efforts.

ADME/Tox

Intestinal absorption

The effect of physicochemical properties on transport processes is complex and interrelated[66]. Consequently, drugs must possess an effective balance between water solubility and lipophilicity[15]. To efficiently evaluate this balance earlier in the drug discovery cycle, many pharmaceutical companies have adopted low- to medium-throughput *in vitro* assays that measure both solubility and membrane permeability. These assays have provided a growing body of data with which to construct computational models.

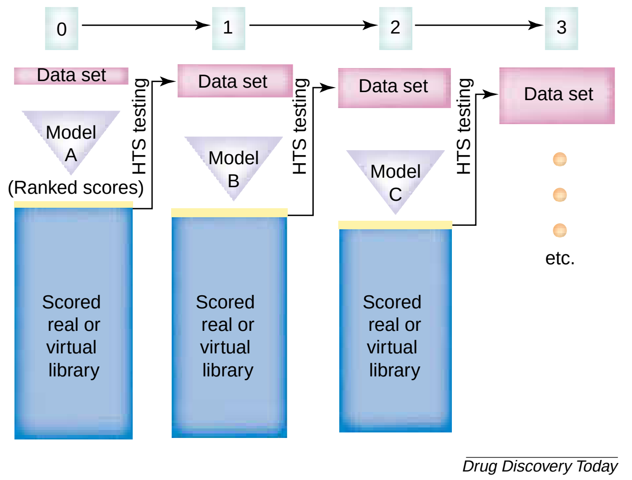


Figure 2. Targeted sequential screening cycles. An iterative (or 'targeted sequential') approach to modeling, virtual screening, prioritization and real screening in HTS.

Solubility Several methods for predicting aqueous solubility based on molecular structure have been reported. Duffy and Jorgensen[67] have recently developed QikProp [Version 1.6 (2001), Schrödinger, New York, NY, USA], a software package that estimates several physical properties, including free energies of solvation in water, octanol, and hexadecane. One of the attractive features of QikProp is the intuitive, physical nature of its descriptors, such as dipole moment, solvent accessible surface area and the number of rotatable bonds. Jurs and Mitchell[68] have developed a neural network solubility model based on a training set of 300 diverse organic compounds. A genetic algorithm was used to choose nine descriptors from a much larger set, including partial positive surface area, charge on the most negative atom and the cube root of gravitational index. However, both QikProp and the Jurs and Mitchell model require the generation of 3D structures. Viswanadhan and colleagues[69] have developed fragment-additive solubility models with partial least squares (PLS) using a combination of simple substructure fragments and UNITY™ (Tripos, St Louis, MO, USA) fingerprints.

Most of the solubility models that have been used[70], including those discussed previously, have achieved very impressive R^2 values on test sets of diverse compounds with data values ranging over ten log units. [See also Syracuse Research Corporation's solubility prediction software:

http://esc.syrres.com/interkow/estsoft.htm; Advanced Chemical Development (ACD) solubility prediction software: http://www.acdlabs.com/products/phys_chem_lab/aqsol/]. A significant challenge that remains, however, is the ability to predict solubility accurately within the narrower, but more pharmaceutically relevant, range of up to 100 µg ml^{-1} (Ref. 15).

Absorption Once dissolved, drug compounds must pass through the intestinal wall to reach the bloodstream. The Caco-2 cell line [among others, such as Madin-Darby canine kidney cells (MDCK)] is used widely as a model for intestinal membrane permeability, and there has been much discussion in the literature regarding the sigmoidal correlation between a compound's molecular polar surface area and either Caco-2 permeability[71] or intestinal absorption[72,73]. Compounds with a PSA of >140 Å^2 have a much lower probability of diffusing through the intestinal membrane. However, the sigmoidal correlation plot becomes much more diffuse when this relationship is examined over a larger, more diverse set of drug-like compounds, presumably because many compounds are actively transported or are efflux substrates. Thus, this sigmoidal relationship is most useful in a classification sense.

Several approaches to modeling intestinal absorption have been described. Artursson obtained good results with a PLS model using descriptors such as PSA, polarizability, numbers of hydrogen bond donors and/or acceptors and a data set of ten endothelin receptor antagonists[74]. Jurs and coworkers[75] constructed a neural network model using descriptors similar to those used in their solubility model, also with encouraging results.

Metabolism

After a drug passes through the intestinal wall, it faces a multitude of metabolizing enzymes in the liver. Several attempts have been made to develop computational models predictive of metabolic liability. In particular, de Groot and coworkers[76] and Wrighton and coworkers[77] have constructed 3D pharmacophore models for several of the CyP450 isozymes, including 2D6 and 3A4. However, there are few reports in the literature of generally predictive microsomal or *in vivo* half-life models. One exception is work done by Martinez and colleagues where neural network models were generated using *in vivo* half-life data from 30 antihistamine compounds using four-atom topological descriptors[78]. It might be necessary, in this context, to construct a series of separate models based on half-life data from individual chemical series. During virtual screening, the choice of model would depend on the similarity of the predicted compound to the modeled chemistry.

Selectivity and toxicity

To evaluate selectivity and toxicity, most pharmaceutical companies evaluate lead compounds in several screens. Perhaps two of the most important recent screens involve the HERG (Refs 79,80) and the sodium type-II (Ref. 81) ion channels, which aid in detecting a compound's potential for QT prolongation and general inhibition of ion channel depolarization, respectively. Although reports in the literature describing the impact of computational chemistry on these two channels are scarce, some companies are using internally generated HERG and sodium channel data to construct computational models.

For general toxicity, Accelrys' TopKat® (Accelrys, Princeton, NJ, USA; http://www.accelrys.com), for example, offers a toxicity software package that uses linear discriminant analysis and multiple linear regression models with 2D descriptors to generate toxicology profiles for organic compounds. Vracko and coworkers[82] have developed a neural network based on LD_{50} data from 41 benzene analogs using 3D geometric descriptors, and obtained correlation coefficients between 0.4 and 0.8.

Summary and future directions

Computational chemistry and informatics have had demonstrable impact on synthesis and screening in drug discovery[9]. Large screening libraries are commonplace and HSS efforts routinely provide fast expansions of those libraries. In some cases, previously disconnected HTS and HSS efforts have become integrated with the rest of drug discovery and with each other through informatics infrastructure and computational chemistry-directed virtual screening. Up until only a few years ago, drug discovery was more linear than it is now, in that ADME/Tox and drug-likeness were assessed only after activity optimization of the lead. Computational chemistry, together with high-throughput methods, has contributed to making the process multi-dimensional, in combining the process for lead generation and optimization with assessment of ADME/Tox and drug-likeness (Figs 3 and 4).

However, there is still great opportunity for increased impact. Targeted sequential screening is used more frequently but has not replaced campaign screening. HTS-generated data is not always used to develop models for subsequent, directed screening efforts. Targeted synthesis is prevalent but the integration between virtual library, virtual screening models and HSS could be greatly improved. It is beginning to be recognized that ADME/Tox profiles need to be generated and predicted early and then considered in the selection of lead series and prioritization of project resources, and that the most potent leads are often not the best, compared with less active, but more drug-like, leads[14].

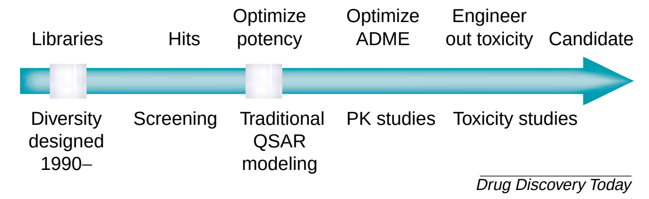


Figure 3. The impact of computational chemistry from 1945 to 1995. A timeline showing the serial nature in which various drug discovery components were brought to bear within a project before 1995. The boxes represent computational chemistry impact. Abbreviations: PK, pharmacokinetics; QSAR, quantitative SAR.

Currently, much of the impact of 3D macromolecular binding-site information, either through structure based drug design (SBDD) or from *de novo* or homology folding techniques, is limited to a minority subset of pharmaceutical targets. Although experimental structure-determination technology will continue to advance, other methods relying on computational modeling will become more important, producing some of the first truly effective uses of these technologies in connection with difficult targets. Sophisticated sequence- and fold-based computational technologies, together with, and guided by, experimental data, will yield binding-site models. Experimental data, such as ligand activity changes resulting from protein mutation studies or ligand affinity labeling information, will be used in an iterative feedback loop with computational methods. These efforts will generate detailed binding site models, which will be more broadly useful in ligand design and prediction.

A major challenge remains regarding non-SBDD approaches. Those that use only 1D and 2D information (topology) and incorporate 3D information only implicitly, if at all, have the major drawback of lacking significant 3D information that could be very important for the SAR.

The alternative non-SBDD approaches use 3D information inherent in pharmacophore patterns or a molecular superimposition hypothesis to presume an alignment for the set of compounds, each in the binding site. These approaches require one or more conformational representations and an alignment hypothesis. Generally, the alignment hypothesis suffers from being based on chemical similarities. It would seem that alignment that is based on biological similarities would be more appropriate; ultimately, what the biological binding site regards as alignment and similarity of compounds is really more significant. Approaches will be developed which incorporate conformational and other 3D information more effectively without forcing acceptance of alignment bias based purely on perceptions of structural similarity. A few attempts have already been described which, to at least some degree, enable the optimization of alignment hypothesis together with compound conformation, both being guided by the biological activity observed[43,83–85]. (See also Catalyst software, Accelrys.)

Important change is anticipated in the kinds of technologies that have been the focus of this review. As the repertoire of computational chemistry and HSS continues to increase, the ability to realize synthesis from computational chemistry directions will improve. High-speed synthesis technologies will become better integrated with medicinal chemistry efforts to form a continuum of synthetic capability. This synthetic continuum will play to the strength of computational modeling to analyze, design and predict populations of compounds with desired project properties, because a larger percentage of predicted and prioritized compounds will be synthesized than is presently the case. The distinctions between medicinal chemistry libraries and HSS libraries will diminish as virtual libraries and real libraries become increasingly drug-like and inclusive of good medicinal chemistry and pharmaceutical development properties and principles. Neurogen, and others, continue to place high value on concepts that have already demonstrated value, including:

- the portfolio approach to modeling, representing multiple descriptor concepts and modeling technologies;
- analysis and predictions of populations of compounds versus individual compounds;

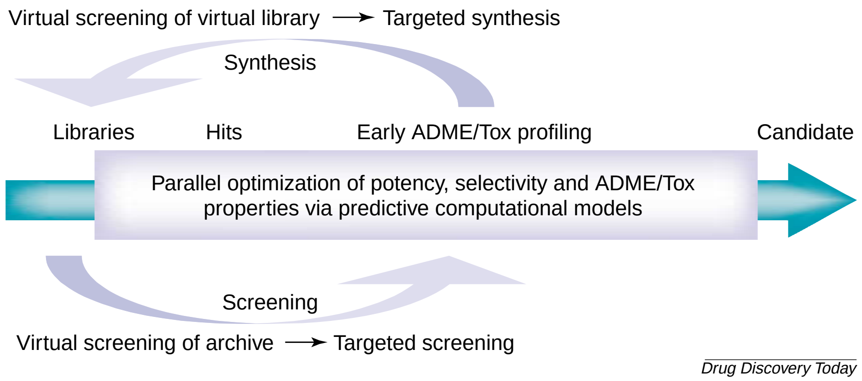

Virtual screening of virtual library
Targeted synthesis
Synthesis
Libraries
Hits
Early ADME/Tox profiling
Candidate
Parallel optimization of potency, selectivity and ADME/Tox properties via predictive computational models
Screening
Virtual screening of archive
Targeted screening
Drug Discovery Today

Figure 4. The impact of computational chemistry from 1995 to 2001. A timeline illustrating the parallel approach to drug discovery adopted during the past five years by many pharmaceutical and drug discovery companies. The rectangle represents computational chemistry impact.

- shifting probability;
- targeted sequential screening and sequential optimization;
- conserved binding motifs and privileged structures for activity-rich libraries and virtual libraries and;
- seamless integration of HSS, HTS and computational chemistry through informatics.

In conclusion, data-mining technologies will be increasingly integrated with traditional computational chemistry. The integrated and iterative use of these methods, together with the substantial data from HTS (including ADME/Tox and drug-like property data), to form models that can be used to prioritize further library screening and synthesis through virtual screening activities, already has significant impact on lead discovery and exploration[86]. Although the road ahead through less validated targets might be difficult[87], the approaches offered by the integrated efforts of informatics, computational chemistry, and SBDD, together with the promise of genomics, makes for an interesting, challenging and fruitful future for modern drug discovery.

Acknowledgements

We thank Dan Severance, Joe Ochterski, Harry Brielmann and Leah Reimer for their proofreading and valuable suggestions.

References

1 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 160–178

2 Morris, G.M. (2001) Mining for medicines – *in silico. Trends Biotechnol.* 19, 123–124

3 Waszkowycs, B. *et al.* (2001) Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Systems J.* 40, 360–376

4 Schneider, G. *et al.* (2001) Integrating virtual screening methods to the quest for novel membrane protein ligands. *Curr. Med. Chem.: Cent. Nerv. Syst. Agents* 1, 99–112

5 Dean, P.M. *et al.* (2001) Industrial-scale genomics-based drug design and discovery. *Trends Biotechnol.* 19, 288–292

6 Leach, A.R. and Hann, M.M. (2000) The *in silico* world of virtual libraries. *Drug Discov. Today* 5, 326–336

7 Hann, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

8 Sadowski, J. (2000) Optimization of chemical libraries by neural networks. *Curr. Opin. Chem. Biol.* 4, 280–282

9 Golebiowski, A. *et al.* (2001) Lead compounds discovered from libraries. *Curr. Opin. Chem. Biol.* 5, 273–284

10 Terstappen, G.C. and Reggiani, A. (2001) *In silico* research in drug discovery. *Trends Pharmacol. Sci.* 22, 23–26

11 Valler, M.J. and Green, D. (2000) Diversity screening versus focused screening in drug discovery. *Drug Discov. Today* 5, 286–293

12 van Drie, J.H. and Lajiness, M.S. (1998) Approaches to virtual library design. *Drug Discov. Today* 3, 274–283

13 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387

14 Lipper, R.A. (1999) E Pluribus Product: How can we optimize selection of drug development candidates from many compounds at an early stage? *Mod. Drug Discov.* 2, 55–60

15 Lipinski, C.A. (2001) Avoiding investment in doomed drugs. Is poor solubility an industry wide problem? *Curr. Drug Discov.* (April) 17–19

16 Tarbit, M.H. and Berman, J. (1998) High-throughput approaches for evaluating absorption, distribution, metabolism and excretion properties of lead compounds. *Curr. Opin. Chem. Biol.* 2, 411–416

17 Smith, D.A. and van de Waterbeemd, H. (1999) Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* 3, 373–378

18 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today* 5, 49–58

19 Eddershaw, P.J. *et al.* (2000) ADMET/PK as part of a rational approach to drug discovery. *Drug Discov. Today* 5, 409–414

20 Johnson, D.E. and Wolfgang, G.H.I. (2000) Predicting human safety: screening and computational approaches. *Drug Discov. Today* 5, 445–454

21 Wang, J. and Ramnarayan, K. (1999) Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* 1, 524–533

22 Ekins, S. *et al.* (2000) Progress in predicting human ADME parameters *in silico. J Pharm. Tox. Meth.* 44, 251–272

23 Lewell, X.Q. *et al.* (1998) RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38, 511–522

24 Bravi, G. *et al.* (2000) PLUMS: a program for the rapid optimization of focused libraries. *J. Chem. Inf. Comput. Sci.* 40, 1441–1448

25 Teague, S.J. *et al.* (1999) The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* 38, 3743–3748

26 Weber, L. *et al.* (1995) Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem., Int. Ed. Engl.* 43, 2280–2282

27 Martin, E.J. and Critchlow, R. (1999) Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* 1, 32–45

28 Ghose, A.K. *et al.* (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* 1, 55–68

29 Villar, H.O. and Koehler, R.T. (2000) Comments on the design of chemical libraries for screening. *Mol. Diversity* 5, 13–24

30 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Deliv. Rev.* 23, 3–29

31 Basak, S.C. *et al.* (2000) Use of statistical and neural net methods in predicting toxicity of chemicals: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* 40, 885–890

32 Xue, L. and Bajorath, J. (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry and virtual screening. *Comb. Chem. High Throughput Screening* 3, 363–372

33 Burden, F.R. and Winkler, D.A. (1999) New QSAR methods applied to structure–activity mapping and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 39, 236–242

34 Egan, J.P. (1975) *Signal detection theory and ROC analysis*. Academic Press

35 DasGupta, B. *et al.* (2000) On computing the nearest neighbor interchange distance. *Discrete Math. Theor. Comput. Sci. Amer. Math. Soc.* 55, 125–143

36 Wilson, D.R. and Martinez, T.R. (1997) Improved heterogeneous distance functions. *JAIR* 6, 1–34

37 Cramer, R.D. *et al.* (1999) Prospective identification of biologically active substructures by topomer similarity searching. *J. Med. Chem.* 42, 3919–3933

38 Andrews, K.M. and Cramer, R.D. (2000) Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* 43, 1723–1740

39 Patankar, S.J. and Jurs, P.C. (2000) Prediction of IC_{50} values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci.* 40, 706–723

40 Pirard, B. and Pickett, S.D. (2000) Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.* 40, 1431–1440

41 Chen, X. *et al.* (1999) Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* 39, 887–896

42 Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis and virtual screening. *J. Chem. Inf. Comput. Sci.* 41, 233–245

43 Hopfinger, A.J. and Duca, J.S. (2000) Extraction of pharmacophore information from high-throughput screens. *Curr. Opin. Biotechnol.* 11, 97–103

44 Guner, O.F. (ed.) (2000) *Pharmacophore perception, development, and use in drug design.* International University Line

45 Friedman, J.H. (1991) Multivariate adaptive regression splines. *Ann. Statist.* 19, 1–141

46 Burnham, R. (2000) Molecular data mining tool: advances in HIV research. *PC AI* 14, 31

47 Steinberg, D. (2001) An alternative to neural nets: multivariate adaptive regression splines (MARS). *PC AI* 15, 38–41

48 Burden, F.R. *et al.* (2000) Use of automatic relevance determination in QSAR studies using bayesian neural networks. *J. Chem. Inf. Comput. Sci.* 40, 1423–1430

49 McGregor, M.J. and Pallai, P.V. (1997) Clustering of large databases of compounds: using the MDL 'Keys' as structural descriptors. *J. Chem. Inf. Comput. Sci.* 37, 443–448

50 Brown, R.D. and Martin, Y.C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* 37, 1–9

51 Martin, Y.C. *et al.* (1998) Validated descriptors for diversity measurements and optimization. *Pharm. Pharmacol. Commun.* 4, 147–152

52 Ghuloum, A.M. *et al.* (1999) Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.* 42, 1739–1748

53 Turner, D.B. and Willett, P. (2000) The EVA spectral descriptor. *Eur. J. Med. Chem.* 35, 367–375

54 Labute, P. (2000) A widely applicable set of descriptors. *J. Mol. Graph. Model* 18, 464–477

55 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

56 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877

57 Osterberg, T. and Norinder, U. (2000) Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics. *J. Chem. Inf. Comput. Sci.* 40, 1408–1411

58 Mason, J.S. *et al.* (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 42, 3251–3264

59 Patchett, A.A. and Nargund, R.P. (2000) In *Annual reports in medicinal chemistry*. (Trainor, G.L., ed.), pp. 289–298, Academic Press

60 Ajay *et al.* (1999) Designing libraries with CNS activity. *J. Med. Chem.* 42, 4942–4951

61 Hann, M.M. *et al.* (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864

62 Leach, A.R. *et al.* (1999) Implementation of a system for reagent selection and library enumeration, profiling and design. *J. Chem. Inf. Comput. Sci.* 39, 1161–1172

63 Kohonen, T. (1997) Self-Organizing Maps. (2nd edn) (H.K.V. Lotsch, ed.), Springer-Verlag, Berlin

64 de Julian-Ortiz, J. (1999) Virtual combinatorial syntheses and computational screening of new potential anti-herpes compounds. *J. Med. Chem.* 42, 3308–3314

65 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

66 van de Waterbeemd, H. *et al.* (2001) Lipophilicity in PK design: methyl, ethyl, futile. *J. Comput.-Aided Mol. Design* 15, 273–286

67 Duffy, E. and Jorgensen, W. (2000) Prediction of properties from simulations: free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* 122, 2878–2888

68 Jurs, P. and Mitchell, B. (1998) Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* 38, 489–496

69 Viswanadhan, V.N. *et al.* (1999) Prediction of solvation free energies of small organic molecules: additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* 39, 405–412

70 Huuskonen, J. *et al.* (1997) Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharm. Sci.* 86, 450–454

71 Luthman, K. *et al.* (1998) Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.* 41, 5382–5392

72 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

73 Luthman, K. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* 14, 568–571

74 Artursson, P. (1999) Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharm. Res.* 16, 1520–1526

75 Jurs, P.C. *et al.* (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* 38, 726–735

76 de Groot, M.J. *et al.* (1999) A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed *N*-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* 42, 4062–4070

77 Wrighton, S.A. *et al.* (1999) Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *J. Pharm. Exp. Ther.* 291, 424–433

78 Martinez, A. *et al.* (2000) Prediction of drug half-life values of antihistamines based on the CODES/neural network model. *Quant. Struct.-Act. Relatsh.* 19, 448–454

79 Crumb, W. and Cavero, I. (1999) QT interval prolongation by non-cardiovascular drugs: issues and solutions for novel drug development. *Pharm. Sci. Technol. Today* 2, 270–280

80 Porsolt, R. *et al.* (1999) QT interval prolongation by noncardiovascular drugs: a proposed assessment strategy. *Drug Devel. Res.* 47, 55–62

81 Catteral, W. (1993) Inhibition of Na$^+$ channels by the novel blocker PD85639. *Molec. Pharmacol.* 43, 949–954

82 Vracko, M. *et al.* (1999) Study of structure–toxicity relationship by a counterpropagation neural network. *Anal. Chim. Acta* 384, 319–332

83 Jain, A.N. *et al.* (1994) Compass: a shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Design* 8, 635–652

84 Jain, A.N. *et al.* (1995) Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT1a receptor. *J. Med. Chem.* 38, 1295–1308

85 Hopfinger, A.J. *et al.* (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* 119, 10509–10524

86 Good, A.C. *et al.* (2000) High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov. Today (HTS Suppl. 5),* S61–S68

87 Rastan, S. (2001) Genomics: saviour or millstone? *Trends Genet.* 17, 247–248