Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Advances in virtual screening

Ingo Muegge*, Scott Oloff
Boehringer Ingelheim Pharmaceuticals Inc., 900 Ridgebury Road, Ridgefield, P.O. Box 368, CT 06877-368, USA

Although the term virtual screening as the in silico analog of high throughput screening has been coined only a decade ago, virtual screening is now a widespread lead identification method in the pharmaceutical industry. A myriad of different methods have been developed exploiting the growing library of target structures and assay data as a basis for finding new lead structures. Exploiting synergies between different methods best utilizes the information available and is at the center of recent developments.

Introduction
Virtual screening has become an integral part of the drug discovery process in recent years [1]. Related to the more general and long pursued concept of database searching [2,3] the term ‘virtual screening’ (VS) is relatively young [4]. Walters et al. define virtual screening as ‘automatically evaluating very large libraries of compounds’ using computer programs [5]. As this definition suggests, VS has largely been a numbers game focusing on questions like how can we filter down the enormous chemical space of >10^60 conceivable compounds [6] to a manageable number that can be synthesized, purchased and tested. Although filtering the entire chemical universe might be a fascinating question, more practical VS scenarios focus on designing/optimizing targeted combinatorial libraries and enriching libraries of available compounds from in-house compound repositories or vendor offerings.

The main goal of a virtual screen is to come up with hits of novel chemical structure that yield a unique pharmacological profile. Thus, success of a virtual screen is defined in terms of finding interesting new scaffolds rather than many hits. Interpretations of VS accuracy should therefore be considered with caution. Low hit rates of interesting scaffolds are clearly preferable over high hit rates of already known scaffolds. Box 1 lists some practical considerations for a VS set-up.

VS has experienced increased attention in recent years (Fig. 1) due to the rise in available datasets, VS techniques and excitement created by successful screening studies. When integrated with high throughput screening (HTS), VS can aid in the rapid identification of novel ligands [7,8]. VS includes target specific search criteria but also target independent considerations such as drug-likeness [9,10]. The use of VS technologies has also aided the identification of bioactive molecules from natural products [11]. VS methods are often divided into structure-based VS (SBVS) [12,13] and ligand-based VS (LBVS) [14,15]. SBVS and LBVS have been reviewed frequently in the literature. Therefore, we focus here on possible synergies between SBVS and LDVS, some new interests in MACHINE LEARNING techniques for VS, and highlight recent success stories.

Synergies between structure-based and ligand-based virtual screening
SBVS and LBVS have been considered almost mutually exclusive suggesting LBVS to be used primarily in the absence of protein target structure(s) and SBVS to be used if target structure(s) are available. Especially when a target protein structure at high atomic resolution is available, SBVS is often considered as the first choice strategy ignoring possible LBVS
alternatives. Until recently, only sporadic studies have pointed to the fact that LBVS offers a strong alternative to SBVS even in the presence of protein structural information. For instance, a comparison of LBVS and SBVS methods using the RECALL of known HIV-1 protease inhibitors as test examples has shown how ligand similarity methods can outperform molecular docking as a VS tool [16]. Recently a more systematic comparative study between SBVS and LBVS involving seven different drug targets for which structural information is available has been published [17]. Figure 2 illustrates that in most cases LBVS techniques measuring compound similarity to known potent molecules outperforms molecular docking, a more computationally intensive SBVS technique that generates and scores putative protein–ligand complexes according to their calculated binding affinities [18]. Similar results have recently been reported in the field of GPCRs using homology models for SBVS and 2D-QSAR, DECISION TREES, and PLS as LBVS methods [19]. Although studies about SCAFFOLD HOPPING through SBVS are scarce [20] it is recognized that SBVS techniques have a better potential to identify compounds with a novel core scaffold. Therefore, SBVS and LBVS should not be applied independently but rather in concert to increase the chances of finding novel hits [21,22]. Some software approaches such as SDocker have begun integrating both strategies by including ligand similarity as a part of the SCORING FUNCTIONS used in the docking algorithm [23]. Bologa et al. has also reported the integration of both strategies to aid in the identification of the first selective GPR30 agonist [24].

Machine learning algorithms in virtual screening
An LBVS approach that is quickly gaining popularity in VS, called machine learning, builds predictive compound activity models that are based on available assay data. Machine learning approaches have been reviewed less before in the context of VS. Therefore, we include here a short synopsis of the most prominent techniques in addition to highlighting their applications in VS. Several recent success stories have been reported in the literature and a few examples of the more
popular machine learning techniques are listed in Table 1. Each method has its own advantages and disadvantages that should be understood to select the best approach for a particular LBVS. The first approach listed in the table, a self-organizing map (SOM), is quite simple and easy to visualize. A dataset of compounds are mapped on to a 2D grid such that most similar compounds are grouped together. Compounds found in the same vicinity of those with a desired biological property are considered potential hits in a virtual screen; however the SOM approach in general has a very high false positive rate. It is a simple and easy way to visualize if compounds that have the same biological properties group together. SOMs have been used recently to identify several purinergic receptor agonists [25]. Another approach called Binary QSAR uses all compounds in the training set to predict the biological property of test compound(s) in a virtual screen rather than just the most similar training compounds. Although not providing an image of the training data, this approach is fast and like SOMs it works well if the training data is highly similar to the test compounds being screened. Compounds that are significantly different from the training set are not expected to be predicted accurately and are commonly missed in a virtual screen [26].

Other approaches require a pre-built model to perform a virtual screen. These techniques form correlations between training data and descriptors that describe training compounds to predict a biological property for a virtual set of compounds. The Bayesian Classifier requires a pre-built model and is somewhat similar to Binary QSAR except it identifies specific descriptors that best distinguish compounds with a desired biological property from others. This search for pertinent descriptors is called variable selection and helps the model eliminate descriptors that are not relevant to the current problem and cloud the separation of one biological property class from another. The Bayesian Classifier algorithm has been shown to not perform as well as even more sophisticated approaches [27]; however, it does handle large training sets much easier.

Decision trees (or Forests) incorporate the simplest form of variable selection and can be considered as a set of Boolean functions. Descriptors that capture molecular features of the training compounds are systematically added to a Decision tree model one at a time until compounds with different biological properties are adequately separated. This approach allows the researcher to easily determine the chemical features most relevant to the target biological property. This information can be used in the design of future molecules. Virtual screening with Decision trees is quite easy as well. Comparison studies have shown it slightly outperforms methods such as the Bayesian Classifier however other more advanced approaches show higher enrichment rates in a virtual screen [27].

In yet another class of machine learning approaches, mathematical function(s) are used to correlate descriptor values with a biological property. The simplest of these builds a linear correlation and is called multiple linear regression. A very popular extension called partial least squares (PLS) helps simplify the model optimization so larger training sets can be easily used. Variable selection techniques are also commonly added above the PLS algorithm to optimize the descriptors used in the linear model. This approach has shown success enriching a virtual database for various GPCR ligands [19]; however, it has an obvious drawback. There is not always a linear correlation between the property modeled and the descriptors describing your dataset. The artificial neural network (ANN) and support vector machine (SVM) approaches allow one to build non-linear correlations. ANN and SVM have become popular tools for model building and virtual screening. In a side-by-side comparison both ANN and SVM show similar enrichment rates for several virtual screens [27].
The k-nearest neighbors approach (kNN) does not require the use of a mathematical function to split one property class from another, which can be very useful when the problem is complex. Compounds in a virtual screen are predicted based on known activities of the most similar training compounds. Similarity between compounds is calculated using only a small set of most pertinent descriptors that are optimized during model building (this optimization can be quite slow for large datasets). When used in a virtual screen kNN shows database enrichment rates similar to both ANN and SVM [27]. A domain of applicability can be used in virtual screening to improve the enrichment rates in a virtual screen by only allowing the model to predict compounds that have the highest chance of being predicted correctly. Applicability domain techniques have been applied very successfully for both kNN [28] and SVM [29]. However, limitations lie in the

| Method used | Protein target | Identified hits |
|-------------|----------------|-----------------|
| Multiple linear regression | CCR5 | Several new derivatives of active molecules were proposed [43] |
| Pharmacophore modeling | Fetal hemoglobin | Novel inhibitors were identified from a large chemical database [44] |
| Consensus scoring using multiple docking approaches | CK2 | Identified a highly potent inhibitor from a chemical database [45] |
| rDock | Chk1 | 10 Novel inhibitors were identified with 9 different scaffolds [46] |
| Catalyst | PPARγ | 2 Partial agonists were found among a large chemical database and validated in vivo [47] |
| Pharmacophore modeling and FlexX docking | GSK-3 | 9 New inhibitors were predicted from three large chemical databases [48] |
fact that a very narrow applicability domain only identifies potential compound hits that are highly similar to the training set compounds. Such hits could have possibly been found using a simpler, less time consuming approach also.

A new forum has been created, the comparative evaluation of prediction algorithms (CoEPrA), which compares how predictive different machine learning approaches are for blind test cases. This forum illustrates which techniques consistently work best and should be an interesting media for testing new machine learning approaches. A link to CoEPrA can be found together with other links of interest in the Links section.

**Recent successes of virtual screening**

A few examples of recent VS applications are highlighted below. In addition, Table 2 shows a collection of reports of recent successful virtual screens.

**Structure-based virtual screening**

Gold [30] docking and subsequent scoring with the PMF scoring function [31] has identified novel inhibitors of the potential cancer target erythropoietin-producing hepato cellular B2 receptor tyrosine kinase domain with measured $K_d$ of 3.3 μM. Docking and scoring results have been combined with pharmacophore modeling aspects and ‘high content’ wet screening techniques using affinity chromatography [32].

Structure-based virtual screening against the target dipeptidyl peptidase IV (DPPIV) has identified chemical starting points for medicinal chemistry follow-up. Docking of compound collections pre-filtered by physical property and medicinal chemistry considerations as well as matching pharmacophores to known DPPIV inhibitors has resulted in 51 compounds with activities between 30% and 82% at 30 μM concentration in an enzyme inhibition assay [33].

Through the combination of homology modeling and docking methods, several successful VS applications have been published recently. They include the discovery of novel lipooxygenase inhibitors [34] as well as a novel cannabinoid CB2 receptor agonist [35].

Combining SBVS and LBVS techniques has resulted in the discovery of a novel family of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) protease inhibitors [36]. Gold docking and mapping CoMFA/CoMSIA models onto the protein active site have been employed to screen through the Maybridge database of 59,363 compounds in search for novel hits. Twenty-one compounds tested have exhibited inhibition below 30 μM IC₅₀. By following up with analog searching through other databases, an additional 25 inhibitors could be identified. This example illustrates the iterative nature of virtual screening. Although a first VS iteration identifies novel classes of actives but does not necessarily contain the most potent compound within a given class, the second iteration focusing on analogs of the newly found class often leads to more potent hits.

**Ligand-based virtual screening**

Recall experiments using SVMs trained on known cyclooxygenase 2 and thrombin inhibitors have been reported recently. Topological pharmacophore-point triangles have been used as molecular descriptors. In a validation study, 50–90% of the known active compounds could be recalled within the first 0.1% of the ranked databases containing the known actives and a list of arbitrary screening compounds. Following on this positive validation, a subsequent VS study has identified several potential Cox-2 inhibitors that have been tested in a cellular activity assay. A newly found benzimidazole derivative has exhibited significant inhibitory activity better than that of Celecoxib [37].

QSAR models have been developed to discover new antimalaria agents. Specifically, virtual screens for finding Ras farnesyltransferase inhibitors with antimalarial activity have been reported. Following successful recall experiments of known inhibitors QSAR models have been used to identify previously unknown antimalariaals. A new arylaminomethylenediamine has been found through VS with antimalarial activity [38].

Ligand-derived pharmacophore models in concert with cell-based activity assays have been used to discover selective

| Table 3. Comparison of VS approaches working in concert |
|--------------------------------------------------------|
| **Method** | **Structure-based & pharmacophore query** | **Structure-based & machine learning** | **Pharmacophores & machine learning** |
| Specific examples | 51 DPP IV inhibitors were identified in a VS using a pharmacophore and docking filter | New SHBG ligands were found using a combined 2D-QSAR and docking filter | New COX-2 inhibitors were found using SVM’s with pharmacophoric descriptors |
| Pros | Not dependent on scaffolds within the training set | Works well for compounds similar to the training set | Very fast for screening large databases |
| Cons | Reduced accuracy for training set-like compounds compared to ligand-based methods | VS hits are biased towards training set-like compounds | Limited scaffold hopping ability |
| References | [33] | [49] | [37] |
11beta-hydroxysteroid dehydrogenase (11beta-HSD) inhibitors shown to block subsequent cortisol-dependent activation of glucocorticoid receptors [39].

An extension to the feature tree approach [40] called MTree has been reported [41]. Here topological molecular graphs of several ligands are combined to a common feature tree that allows for matching corresponding functional groups. These functional groups are derived akin to pharmacophore queries from a set of diverse but active ligands against a given target protein. Applying this new multiple feature tree approach to recall experiments of known angiotensin converting enzyme inhibitors and a1a receptor antagonists has led to significant enrichments of known active compounds validating the concept of MTrees.

Conclusions
The prevailing opinion has been for a long time that in the presence of a high-resolution target protein structure one should use SBVS whereas in cases where only ligand information is known LBVS should be used. Recent publications have somewhat challenged this view focusing on using LBVS even in the presence of target structure information. Advancements in VS have therefore been made in understanding the strengths and weaknesses of existing methods and how to use them rather than coming up with new approaches. The majority of reported VS successes make the best use of several informational sources. For instance, pharmacophore models using known ligands are combined with homology models; QSAR models are combined with docking approaches. Using all available information in concert is essential for obtaining optimal VS results making each VS experiment unique. Consequently, an increasing number of successful VS applications use more than one VS technique.

Machine learning techniques have been increasingly applied to virtual screening. This is not surprising as ligand-based virtual screening approaches have experienced growth in general through expansions in available chemical libraries, published compound assay data, and the surge of new molecular descriptors and techniques used in similarity comparisons. As it becomes available, experimental data is incorporated into models that are used to aid the design of new compounds. This helps to reduce redundant compounds from being synthesized and to identify molecular features that are important for the biological property of interest.

Although VS has been in a race with experimental high throughput screening techniques in the past decade to increase the speed of processing more and more compounds in less time, this race has slowed in recent years. Some companies such as Aventis have decided to limit the number of compounds to be tested in high throughput screens to increase the quality of hits [42]. Likewise, the focus of VS is now on increasing reliability, hit rates, and the number and quality of novel scaffolds to be discovered rather than speed. Several future challenges are highlighted in the Outstanding issues box.

Several of the most recent virtual screens have resulted from a combination of different approaches utilizing multiple sources of information rather than just structural or ligand assay data (examples in Table 3). Such a combined approach uses ligand-based methods to identify compounds with features important for the target property. Structure-based techniques are used to ensure that the shape, size and energetic interaction potential of the putative ligands complement that of the target protein. Recent successes illustrate the advantage of utilizing all available information in concert making each VS experiment a unique endeavor.

References
1 Stahl, M. et al. (2006) Integrating molecular design resources within modern drug discovery research: the Roche experience. Drug Discov. Today 11, 326–333
2 Martin, Y.C. et al. (1990) Searching databases of three-dimensional structures. In Reviews in Computational Chemistry, (Vol. 1) (Lipkowitz, K.B. and Boyd, D.B., eds) pp. 213–263, VCH Publishers
3 Good, A.C. and Mason, J.S. (1995) Three dimensional structure database searches. In Reviews in Computational Chemistry (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 67–117, VCH
4 Burns, R.F. et al. (1995) Virtual screening as a tool for evaluating chemical libraries. Exploiting Molecular Diversity, Symposium Proceedings 2, Cambridge Healthtech Institute
