Computational Chemistry on a Budget – Supporting Drug Discovery with Limited Resources

Henriëtte Willems[1†], Stephane De Cesco[2†], Fredrik Svensson[3†]*

1. The ALBORADA Drug Discovery Institute, University of Cambridge, Island Research Building, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0AH

2. Alzheimer's Research UK Oxford Drug Discovery Institute, NDM Research Building, Old Road Campus, University of Oxford, Roosevelt Drive, Oxford, OX3 7FZ

3. Alzheimer's Research UK UCL Drug Discovery Institute, The Cruciform Building, University College London, Gower Street, London, WC1E 6BT

*corresponding author: f.svensson@ucl.ac.uk

† These authors contributed equally to this work

Abstract

An increasing number of new drugs have their origin in small biotech or academia. In contrast to big pharma, these environments are often more limited in terms of resources and this necessitates different approaches to the drug discovery process. In this perspective, we outline how computational methods can help advance drug discovery in a setting with more limited resources and we share what, based on our experience, are the best practices for these methods.

Introduction

Government and charity funding of academic research has been shown to have a major impact on drug discovery by elucidating disease biology and de-risking targets.[1] An increasing number of academic contributions to new drugs are made through dedicated academic drug discovery institutes that aim to translate basic research to proof-of-concept.[2–5] In parallel, a growing number of new drugs come from small biotech companies rather than big pharma.[6]

Predictive modeling and informatics are today cornerstones of drug discovery.[7] Computational methods can have an impact from the first conception of a drug discovery project all the way up to clinical trials.[8] Data mining and analysis approaches can help to better inform and greatly speed up the process of target assessment.[9] Virtual screening[10] (VS) is a well-established computational method that is used to find hits for selected protein targets. Docking, QSAR analysis, and matched molecular pairs (MMP) support medicinal chemistry programs to turn hits into leads. More recent additions to the computational toolbox include big data analysis as well as artificial intelligence methods (usually in the form of deep neural networks).[11] Together, computational methods have played an important role in the discovery of several drug candidates and approved drugs.[12]

Academic drug discovery centers and smaller biotech companies often do not have all the capabilities of large pharma, placing certain constraints on what tools and data can be accessed. However, computational methods accessible to everyone can speed-up and reduce the cost of the drug discovery process in a number of ways. In this perspective, we outline the challenges and opportunities for computational methods to impact drug discovery in the context of a resource limited drug discovery organization. We hope this can serve to illustrate the value of these methods across the drug discovery spectrum and that we can help introduce these methods also to non-experts who are curious about what their organisation could gain from computational approaches.

Impact of computational methods on target identification and validation

The first step of most drug discovery programs is to identify and, as far as possible, validate a suitable target. Informatics can be leveraged to sift through large amounts of data to help in this

endeavor.[9] Choosing the right target to start a drug discovery program has never been an easy task. The amount of information available nowadays has the potential to make this decision more informed. It is worth mentioning that the purpose of this perspective is not to provide the "right" approach to select a successful drug target but rather underline the contribution computational chemistry can make and the challenges that a computational scientist will face in this endeavor.

The vast amount and variety of data that is accessible to researchers makes the target selection and validation a discipline by itself. With data comprising CRISPR-Cas9 screens,[13] protein expression profiles, biomarkers, multi-omics studies, and patient data; juggling between different metrics, ontologies, and conventions is required to extract information that can be used to infer relevance in a disease of interest. To help with this task, a plethora of tools exist (see Table 1). We are focusing only on a few that are particularly accessible and provide both a user-friendly interface and access to various data sources. Initiatives such as Open-Targets[14], UniProt[15], and ChEMBL[16] (Table 1) provide an extremely useful starting point to cover areas such as disease association, protein annotation, and potential ligands, respectively. These tools require little computational expertise to operate, and the output parameters are generally well-documented. Often, these portals will be used to gain knowledge of a potential target and build up a picture of the amount and type of information available. Reading the literature linked to the information in the portal helps to further validate or invalidate a target hypothesis. While this approach has its merits, it falls short when the validity or tractability of hundreds of potential targets coming from a genome-wide association study or multi-omics analysis needs to be assessed. It is in this context that informatics can play a role in the integration of all the available resources in an automated and standardized manner.

Whether originating from a genetic screen or an interest in a particular protein family, the list of potential protein targets to investigate can be long. For all these targets data will need to be extracted and combined from multiple sources. At this stage of the project, technical skills such as scripting (e.g. Python or R) and database extraction (e.g. SQL) are important in order to manipulate the information that might come out in different formats. It is common for publicly available data to

be distributed in the form of a database, flat comma-separated files, or as an API (application programming interface) that can be directly accessed by scripting languages. In addition to technical skills, the ability to understand and critically assess both the quality and relevance of the data gathered is essential. This is often a challenge in smaller settings where a specialist in each of these areas is generally not available. As a result, significant effort is required to interpret and analyze the breadth of information available on targets and summarize the information in an actionable manner. The diversity of the data makes it difficult to aggregate and normalize it in order to build metrics for target selection. Another challenge is the sparsity of the data obtained. For example, how can two targets be compared if there is little to no overlap between the data sources available for them? Predictive models that attempt to fill these gaps could offer a solution, but their application often requires specialized knowledge.[17] Also, the multiplication of sources makes it more difficult to keep everything up to date as it requires tracking and going back to each source to check for novel information.

Target selection in practice. When looking for information on a single target, the Open-Targets initiative has done a fantastic job at presenting and providing an easy access to different data sources in a single place. While we acknowledge that a platform like Open-Targets is very well suited for gathering information on a single target, it is more difficult to interpret or extract the information for a bigger list of targets. It highlights the challenge to directly compare different targets, given the spread and fragmentation of the data available. In our institutes, we decided to build a tool, TargetDB (https://github.com/sdecesco/targetDB), to help in this task. The aim of the project was to develop a tool that can collect standardized information on a target of interest into a single file and can be used to prioritize a list of targets by a user-defined score. Data are collected from the above-mentioned resources as well as others, and a series of data analyses are performed in order to extract the most relevant pieces of information for target tractability assessment; a schematic of the process is provided in Figure 1a. Recently, ML algorithms have been applied to target identification and drug discovery in general.[18] It is important to note that these algorithms need well-curated,

uniform and standardized data to maximize their predictive power. We believe initiatives like Open-Targets or TargetDB play an important role in providing data to improve prediction made by these algorithms. In our institutes, TargetDB was used to rapidly prioritize and select targets from an entire family of proteins with the help of a machine-learning (ML) model that classifies targets into three tractability classes (Tractable, Challenging, Intractable) (Figure 1b).

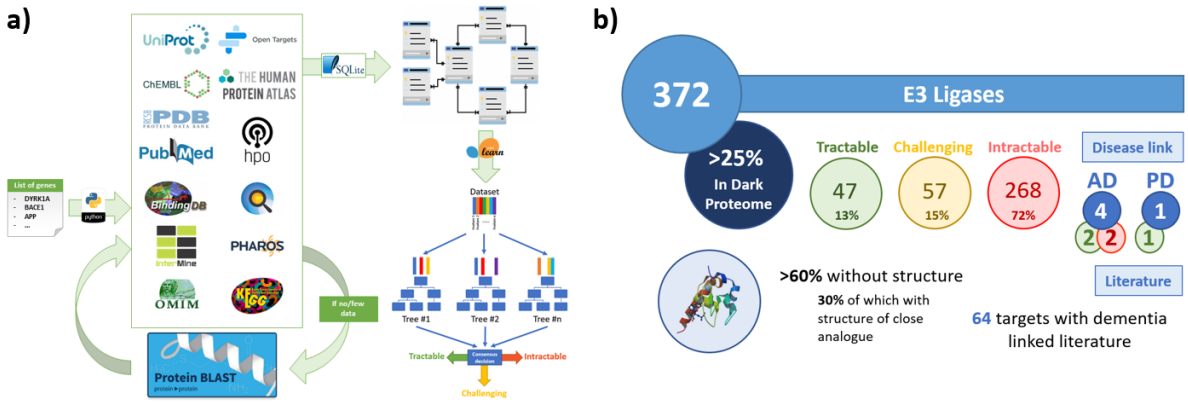

Figure 1. a) Schematic showing how TargetDB searches for information in a range of databases and compiles the results to generate the output **b)** Example of analysis performed on the E3 ligases family using TargetDB to assess potential tractability and disease relevance of targets.

Impact of computational methods on hit discovery

Once a suitable target or phenotype has been determined, the next step is to identify compounds able to bind or modify the selected target/phenotype that can serve as a starting point for medicinal chemistry.

High throughput screening. High throughput screening (HTS) has long been the go-to method for hit-finding for drug discovery.[19] However, due to the high costs, this method has been out of reach for many academic labs and smaller biotechs. Computational methods can help to make the screening process more manageable, either through VS campaigns or through the application of ML-driven iterative screening and rational library design.

In iterative screening, a subset of the compound library is screened and the results from this screen used to inform on the next stage of screening. Studies have shown that this approach can retrieve most of the active compounds while screening less than half of the total screening library.[20] Typically, ML methods are used to predict the next set of compounds to screen. While this approach makes the screening logistics more complicated by introducing multiple rounds of compound picking, the reduction in the number of compounds to be screened can more than make up for this, especially for complicated and costly screens. With methods such as automated compound dispensing becoming more commonplace, this trade-off is set to become increasingly attractive. Also, these methods can be used to pick compounds iteratively from a vendor library, only purchasing the compounds of interest.

For many smaller research outfits, the work and cost involved in maintaining a large HTS library are prohibitive and one might look to either maintain a smaller library or to purchase a set of screening-ready plates for each assay (many of the suppliers discussed in the VS section also provide larger libraries in bespoke formats). In either case, it is important not to include compounds that are unlikely to lead to productive starting points for medicinal chemistry.[21] This calls for a rationally designed screening library. Library design can be done either for a specific target, by trying to enrich relevant chemotypes specific for that target, or for a library intended to be screened against multiple different targets.[22,23] Common tasks include filtering of reactive[24] and interfering groups (such as PAINS[25,26]) as well as controlling key molecular properties and chemical diversity. Typically such property filtering is inspired by the concept of lead-likeness, looking for compounds that after development will still end up within drug-like space.[27] Suitable cut-offs for these properties have been reviewed extensively elsewhere.[21] There are several free tools, such as RDKit and Knime, which can be used for compound filtering (Table 2). Substructure filters for PAINS and reactive groups can be downloaded or created in Knime/RDKit. ChemAxon[28] also offers free tools to academics to do this, including logD and pKa calculators (not available in Knime or RDKit). The latter two properties

are needed to calculate a CNS MPO score,[29] which prioritizes ligands likely to penetrate the blood-brain barrier and is therefore an important filter for CNS projects.

For HTS hits, potency often increases with molecular weight, but the most potent molecule might not be the most tractable. Generally, starting with a smaller molecule is desirable.[30] We find that ligand efficiency metrics such as Ligand Efficiency (LE) and Lipophilic Ligand Efficiency (LLE) are useful in prioritizing hits.[31]

Virtual Screening. Virtual screening[10] (VS) refers to the use of computational tools to select compounds for screening in biochemical assays. This is often a key task for a computational chemist and thus we have dedicated a substantial part of this Perspective to discuss what we believe to be the best practice for this task. The term VS is often used to refer to the docking[32] of large compound databases, but there are several alternative techniques such as shape and pharmacophore searching that can also be used to virtually screen compound databases.[33] Screening a set of compounds selected by VS is typically much cheaper than running a high-throughput screening (HTS) campaign because both the compound cost and the cost of screening consumables is lower. It can also be significantly faster and equally or more successful than HTS.[34,35] Our experience is that VS costs around 10-fold less than HTS and takes about half the time. Academic and small biotech drug discovery teams should therefore consider if VS is an option for their projects.

Suitable targets for VS. The targets of many drug discovery programs are not well explored, do not have a crystal structure, and/or do not have many or even any known ligands. VS is challenging in these scenarios, but often, it is still possible to pursue a VS campaign successfully. The key is to find all available information about the target structure and ligands before deciding on a protocol.

Protein structural information can be found in the PDB (Table 1). If there is no structure for the target in the PDB, a BLAST search with the target sequence in UniProt (Table 1) with the PDB as the target database may reveal homologous proteins with crystal structures. Any protein with >25% homology in the relevant domain, e.g. the ATP-binding or protease domain, may yield a useful homology model.[36] Lower sequence homology does not necessarily decrease the chance of success

in VS, but there is a weak correlation between sequence identity and VS enrichment.[36] An available protein structure is a great start for a VS campaign, but protein structures are not all equally useful.[37] Structures with drug-like ligands give a better chance of success than structures with native ligands or substrates, because the best VS enrichments are generally obtained with protein structures whose bound ligands are similar to the compounds to be docked.[38] Potent ligands have a higher likelihood of success, because they typically make more and stronger interactions, and this information can be used to guide the VS. Structures without ligands have less chance of VS success, because the wrong pocket may be targeted, or structural changes in the protein may occur upon binding.[35,39] Good resolution (<3.5Å), and a well-defined active site with residues fully visible in the density, are also useful indicators of the likelihood of success.[37] Table 3 summarizes the hierarchy of desirable features for a VS starting point.

Table 3. Features of VS starting points ordered by likelihood of success. The color gradient highlights the likelihood of success in different scenarios, with green indicating a higher likelihood and red a lower.

Structure-based	Multiple chemotypes	Single drug-like molecule	Affinity >1μM	Ligand not drug-like	No ligand
Multiple structures					
<3.5Å/complete density in active site					
>3.5Å/missing density					
Homology model ensemble					
Single homology model					
Ligand-based	Multiple chemotypes	Multiple Analogues	Single rigid ligand	Single flexible ligand	Affinity > 100 nM

It is always worthwhile retrieving the electron density map from the PDBe (Table 1) to check how well-defined the ligand and pocket residues are.[40] This information can be used to fine-tune the size of pharmacophore constraints or to allow flexibility for certain residues in a docking protocol.

Databases such as ChEMBL[16], Probes&Drugs[41], and PubChem[42,43] (Table 1) can be used to find known ligands. Patents can be mined for ligands using SureChEMBL[44] (Table 1). If ligands, or even a single ligand, for the protein target are known, VS based on ligand shape and pharmacophoric features can be tried. These approaches do not require a 3D protein structure, though it can be used if available. A number of commercial and academic packages are available for ligand-based VS. We have had successful screening campaigns with ROCS, Blaze, MOE, and Phase (Table 2). Ligand-based 3D pharmacophore approaches assume that all the features of the known ligand are important for binding (though this can be manually overridden in some packages).[45] Ligands with only the pharmacophoric features required for potency therefore work best as queries.[45] Conformational flexibility adds complexity to ligand-based 3D pharmacophore VS.[46] More rigid query molecules should therefore be chosen over more flexible alternatives if possible.[46,47] However, not all ligand-based screening tools are sensitive to the query conformation.[45,48] If multiple ligands are available, an alignment can indicate the key binding features and likely pocket shape. For both structure-based and ligand-based screening, data on inactive ligands are also useful to test if the VS protocol is predictive and can differentiate between actives and inactives.[45]

Selecting a database to screen. The decision of which database to screen is a significant factor in the success of a VS campaign. To have a timely impact on a drug discovery program, compounds selected by a VS protocol need to be affordable, deliverable in a reasonable timeframe, and in a suitable format. When considering the cost of a VS, it is important to consider how many compounds should be purchased. Three or four small clusters of actives would be a good outcome of a virtual screen, as this allows for some attrition due to flat SAR, intractable chemistry, or ADME properties that cannot be optimized without losing potency. These are all reasons why we discontinued chemistry on screening hits. For representative examples of our successful VS

campaigns, the hit rate has been 0.5-1.5 % using an IC_{50} < 10 µM in an ADP Glo assay as a cut-off. It is hard to know whether this is typical because hit rates reported in literature use a wide range of cut-offs and different targets have varying rates of success.[30,33] Also, many of the studies that report higher hit rates are retrospective studies that use databases seeded with known actives. These studies have a much higher ratio of actives than typically found in HTS campaign (around 0.05% with some variation for different target classes).[30,49] A report from the Shoichet group comes to similar conclusions.[50] This means that to find 3 or 4 small clusters, so 6 to 10 hits, around 1000 compounds need to be purchased and tested. The cost of buying this many compounds is likely to limit the vendors from which compounds can be sourced and should therefore be considered before embarking on a VS campaign.

Costs vary from approximately $2 to $120 per screening compound, depending on the vendor, quantity required, and the number of compounds in the order. Because the cost per compound typically drops when more than a threshold number of compounds are ordered, a limited budget often goes further when compounds are ordered from a single vendor. Using a single vendor has the additional advantage that ordering and processing the physical compounds is easier and shipping costs are lower. We therefore recommend screening single-source vendor databases before aggregated compound collections (Table 4). Table 4 show some compound vendors and databases, this list is by no means exhaustive but contains the single source vendors we have experience with and whose cost per compound was ≤ $10 for orders over 1000 compounds when we enquired.

Table 4. Some compound databases for VS. Single source vendors included have quoted ≤ $10 per compound for 1000+ compounds purchased; this is not an exhaustive list.

Name	Link	Single source/aggregate	Database size	Purchase from site
BioAscent	compoundcloud.bioascent.com	Single source	125,000	y
ChemBridge	chembridge.com	Single source	1,300,000	y

ChemDiv	chemistryondemand.com	Single source	1,500,000	y
Enamine	www.enaminestore.com	Single source	3,500,000*	y
eMolecules	Emolecules.com	aggregate	>7,000,000	y
MolPort	molport.com	aggregate	>7,600,000	y
Zinc	zinc15.docking.org	aggregate	>230,000,000	n

*3,500,000 compounds available in stock at low price bracket. 1.2 billion compounds available on demand from the Enamine REAL database at higher cost.

In addition to cost, the content of vendors' libraries is also an important factor to consider. Some vendors may simply have a lot more examples of a chemotype of interest than others. So, if the property space and features of the compounds to be purchased can be defined in detail, it may be worthwhile mining very large compound databases for molecules that match the requirements very closely. We recently obtained a hit rate of 5% (IC_{50} < 10 μM) from 111 purchased compounds tested in a biophysical screen against a target that had not yielded any hits from a previous HTS. The 111 compounds were selected based on presence of a novel chemotype and fit to a docking model fine-tuned to discriminate between around 100 internally tested actives and inactives. 50,000 molecules with some similarity to known actives were selected for docking from the Enamine REAL database (Table 4) using the infiniSee software from BioSolveIT (Table 2).

A final consideration in database selection is perhaps whether the compound set is manageable, a database of a million or so virtual compounds can be processed easily on a workstation with multiple cores and docking a set of this size may take a weekend. Beyond that size of database, significant time will be required to set up computational infrastructure and workflows.

Preparing a database and search query for VS. To ensure that only suitable starting points for a chemistry campaign are screened, the PAINS, reactive groups, and property filters discussed above

for rational design of screening libraries should be applied before VS. Docking and most ligand-based screening applications need input ligands to be represented in all their likely forms, including charged states, tautomers, and stereoisomers in 3D. Many commercial and free software packages have tools to do this (see Table 2), but the results they deliver and the time they take vary. For us, MolConvert from ChemAxon to generate charges, tautomers, and stereoisomers, followed by geometry optimization in RDKit worked very well without tying up software licenses used for other applications.

Not only the database, but also the search query, be that a protein active site, pharmacophore or ligand need careful preparation. Proteins need to be correctly charged and protonated so that relevant hydrogen bonding and charge interactions can be found.[51] Water molecules need to be assessed and decisions made on whether to keep or remove them. If this is unclear, or where side chains or loops are flexible, the best approach may be to use multiple protein models for the virtual screen.[36,45] The steps required to adequately prepare a protein structure for docking have been discussed extensively elsewhere.[37,51] If a ligand is used as a 3D query, it needs to be in a likely conformation. In the absence of a binding model, the lowest energy conformer is typically used, but Kirchmair et al. showed that when using ROCS the query conformation does not impact performance.[48] Low energy conformations can be found by conformational analysis, followed by optimization with a semi-empirical or QM method, and validated by looking at similar ligands in the Cambridge crystallographic database (Conquest, Table 2) if available. Detailed conformational analysis can also be very useful for pharmacophore generation from multiple ligands.[46] Excluding unlikely conformations, e.g. cis-amides, rings with axial substituents, in the pharmacophore generation reduces the number of possible pharmacophores and improves the likelihood of success.[46] All screening queries should be tested first by seeing if they retrieve known actives, and secondly by their ability to discriminate between actives and inactives, if sufficient activity data is available. Scior et al.[45] have written an excellent, and much more detailed discussion of the pitfalls of binding site and pharmacophore preparation, and other limitations of VS.

Selecting docking software. Many different software packages for VS are available. The ones we have used successfully are listed in Table 2, but numerous other good software tools are available. For docking tools, there have been several competitions in which groups using a range of different strategies have gone head to head in predicting the binding poses and ranking of ligands for which the crystal structure has not yet been released (for example: Gaieb et al.[52] and Carlson et al.[53]). Studies that compare VS success have also been published, e.g. Su et al.[54] These are all useful resources when selecting a docking package and strategy. All docking packages have different strengths, so consider the target and library details when choosing. For example, open source docking programs are not limited by licenses, so are great for running on many CPUs. The authors like GOLD for scenarios where water molecules may form key interactions with the ligand, because it can toggle water molecules in the binding site on and off during the screening run. However, we prefer Glide in other cases because it calculates ligand strain energy, which is very helpful when evaluating poses. Yuriev et al.[55] have written an excellent review that discusses which docking tools can handle flexible proteins, solvation and fragments among other things.

Docking packages generally have multiple scoring functions, so the validation of the docking protocol should include assessing the best scoring function. Many VS packages also have a rescore mode, so it is possible to rescore with a scoring function that was not available in the tool used for the docking. Ideally, this should be done by optimizing the docking pose slightly to the new scoring function. The results of multiple scoring functions can then be combined in various ways to improve enrichment. A recent example of this is a report from Ericksen et al.[56] who use ML to improve traditional consensus scoring models. Success has also been reported in combining structure-based and ligand-based methods.[55,57] Ligand-based methods can provide a quick pre-filter to reduce the number of compounds to submit to docking, which is typically slower. Alternatively, ligand-based methods can be used as a post-docking filter to ensure that all docking hits make the required interactions with the receptor.[58] The latter approach has been very successful in our hands in

increasing the enrichment achieved by the VS. Ligand-based methods have also been shown to be very successful on their own.[33,59]

Selecting compounds to buy and test. The final computational step in a VS campaign is deciding which compounds to purchase. This is an important step that potentially has more impact on the success of the VS campaign than, for example, which scoring function is used. Scoring functions are quite poor at ranking compounds,[52,54] so all compounds with a reasonable score (e.g. similar to that of known ligand) should be considered for purchase. A well-known problem with scoring functions is that the score increases with molecular size[60]. This can result in more attractive, smaller compounds being overlooked. Using a ”virtual ligand efficiency” score, by dividing the score by the number of heavy atoms,[30] for example, or dividing the hit list into molecular weight tiers, and picking a set from each one[34] can overcome this issue. These strategies should be combined with a clustering step to ensure diversity. However, picking a few examples from each cluster is useful, because it allows some SAR to emerge.[49] If the set is too diverse, it can be difficult to prioritize what to work on. Data Warrior is a useful tool for this type of clustering, because it clusters by Tanimoto similarity. Similarity of 0.7-0.8 tends to produce clusters of genuinely similar molecules, which is more difficult to achieve with k-means and hierarchical clustering algorithms. It is essential that any hits identified from the initial screen are resupplied as solids or resynthesized in-house to properly quality control the compound and then verify its activity.[49] As mentioned above, we have successfully carried out a number of VS campaigns against various targets. Figure 2 shows a representative VS funnel that was deployed in one such project.

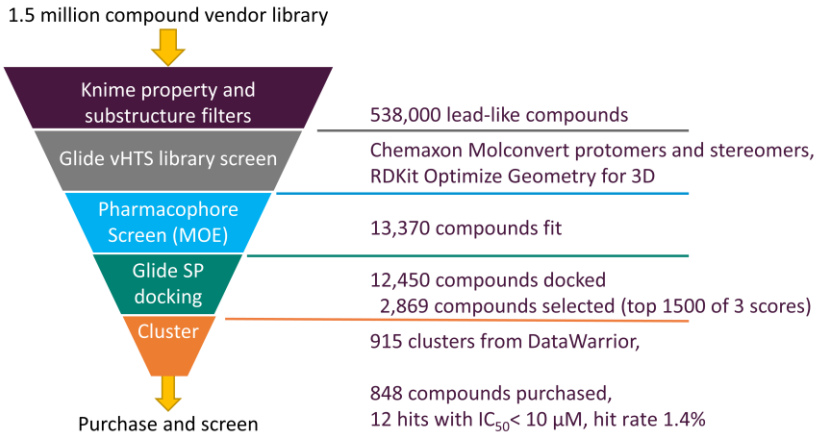

Figure 2. Example of a VS funnel successfully used in-house. The tools used at the various steps are indicated in the figure.

Impact of computational methods on the hit-to-lead stage

Once a suitable starting point has been identified, the next task is to develop it into a lead compound with a good target potency as well as other favorable characteristics. At this stage too, computational methods can speed up the process and increase the quality of the final lead. Docking studies and prediction of ADME properties are useful to guide the design process and can lead to better molecules faster. We routinely employ these methods in-house and Figure 3 show the final compound of one series from our Notum inhibitor discovery program that was optimized with the help of these methods.[61]

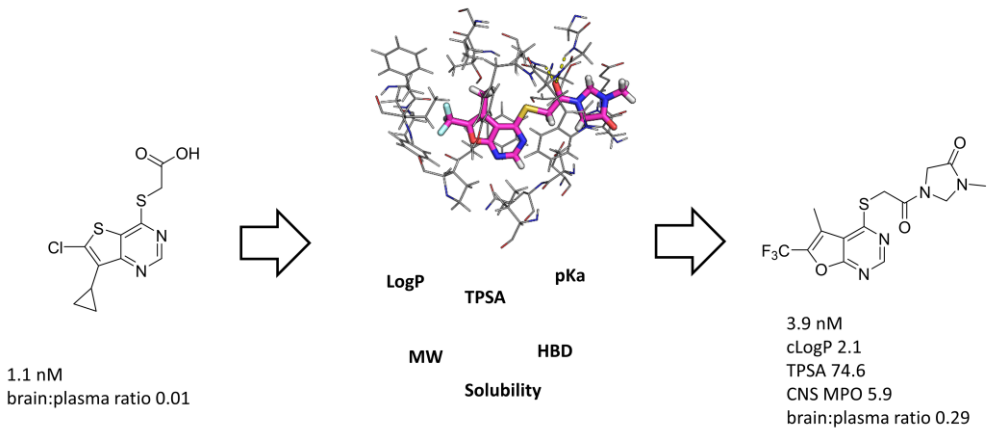

Figure 3. Docking, using Glide (Table 2), and various property predictions were used to guide the development of a series of furanopyrimidine amides as inhibitors of Notum based on a non CNS penentrant lead. The final compound displayed resonable CNS properties.[61]

If a crystal structure is available, the docking methods discussed in the previous section can be applied to generate design ideas and to rank compounds according to predicted binding affinity. In some cases, more accurate but also more computationally intense methods, such as free energy perturbation (FEP) or molecular mechanics generalized Born surface area (MM-GBSA)/molecular mechanics Poisson–Boltzmann surface area (MM-PBSA), have been shown to provide a better correlation with measured affinities and therefore a better basis for compound optimization.[62] However, these methods are not very accessible on a modest budget. They require expensive licenses, a lot of computing time and need to be calibrated with a lot of data, in the authors' experience. Also, their applicability domain tends to be small.[47] Combined with structure-activity relationships established for the series, structure-based optimization can be a powerful tool to quickly generate better compounds.[63]

QSAR modeling. Quantitative structure-activity relationship (QSAR) and Quantitative structure-property relationship (QSPR) models have long been used to inform on compound design.[7] The idea is to create a function predicting the property of interest from the compound structure. These models are often constructed using ML methods and use molecular fingerprints[64] or a set of molecular descriptors to describe the input molecules. QSAR models can be used to prioritize which molecules are most likely to meet the design criteria and can span multiple endpoints, including basic molecular properties, biological activity, and metabolic stability. A challenge with these approaches is that they require data to base the models on, and there may not be enough data available to build a model at the start of a project. For successful QSAR modeling, both negative and positive examples are required. However, basic molecular properties, as well as many ADME-T/DMPK properties, are transferable between projects and basic properties can often be used as proxies for other endpoints. For targets from families where similar proteins have been researched,

there is also the opportunity to use information from these related targets to inform on the target at hand. QSAR models can be purchased pre-trained as part of a software package (e.g. ADME models in StarDrop) or be constructed in-house and trained from available data. Open packages such as caret[65] in R or scikit-learn[66] in Python are commonly used for building ML models and some commercial software packages also provide this feature. Best practices for QSAR modeling has been published elsewhere.[67] While in-house models offer more flexibility, pre-trained commercial models are a convenient choice for groups with limited data or domain experience.

Recently, there has been an increasing interest in the use of deep neural networks for QSAR applications and in many settings, these methods have shown better performance than traditional approaches.[68] However, these methods are generally very computationally expensive and the gains for many tasks are not that large in relation to other methods.[69] Thus, it is our experience that for most standard applications in a small institute, the additional time and hardware investment needed for these methods might not be warranted.

Importantly, QSAR models are most often not intended to replace the experimental assay but to select compounds more likely to have favorable properties prior to synthesis and thus reduce the number of required design cycles. Studies have shown that incorporation of QSAR predictions improves the overall quality of compounds in projects.[7,70]

Matched molecular pairs. The improvement of ADME-PK properties is an important aspect of lead development.[70] As discussed above, QSAR models can be used for ADME-PK modeling but another popular technique is the use of matched molecular pairs (MMP).[71,72] This approach relies on the identification of sets of very similar compound pairs, typically differing only by one chemical transformation, with associated data for the property under investigation. Once a database of such transformations has been established, it can then be used to evaluate potential changes to a lead molecule by looking at the average change in the property for the corresponding changes to molecules in the database. One of the advantages of this technique is that the predictions are readily interpretable and the examples behind the predictions can be reviewed.

While feasible for any property, MMP requires a lot of data to give robust estimates and are therefore most suited for properties that are transferable between projects. MMP can be particularly useful for predicting microsomal stability, efflux, and cytochrome P450 inhibition changes, which are often substructure dependent, and therefore not easy to predict with QSAR methods. ChEMBL is an excellent source for extracting clearance, permeability, and other ADME data which can then be used to build MMP. Several software options are available for matched molecular pair building (Table 2).

Quantum-mechanical (QM) calculations. QM calculations can also be very useful in the hit-to-lead stage. They can be used to identify strain in (putative) bioactive conformations and to develop hypotheses to relieve this. For example, Kuhn et al. report successfully applying QM methods using Guassian98 to relieve the torsion angle strain between two heterocycles,[47] while Heightman et al. use QM-based single point and minimum energy calculations with Q-Chem to optimize interactions between two regions of their ligand that are in close contact.[73] Along with the QM package Jaguar, that we used for more accurate pKa prediction, we have also used the open-source QM package ORCA[74]to calculate the activation energy of the reaction between a nitrile and a cysteine to form a thioimidate covalent bond as exemplified by Cavalli et al.[75]. This allowed us to guide and tune the design of our potential inhibitors with useful information on their reactivity.

Impact of computational methods on lead development

The further the discovery process progresses, the less data for the relevant design stage tends to be available to build predictive models on. For example, while a large set of cell-based data can be obtained quite easily, the number of compounds tested in an animal model will be significantly fewer. In our settings, this is a difference between thousands and a handful of datapoints. Generally, this means that predictive modeling tends to play less of a role in the later stages of a project but there are some areas where computational models still can contribute. Some off-targets and toxicity mechanisms are routinely evaluated using computational models. Probably the most commonly predicted off-target activity is hERG, where good quality models can be obtained.[76] In addition,

carcinogenicity can be reliably evaluated with computational methods.[77] Another area where computational methods can be useful also in the advanced stages of drug discovery is the prediction of metabolites and metabolic stability.[70] For both of these tasks, both commercial and open solutions are available,[78] perhaps the most prominent being the various tools offered by Lhasa Limited. It is important to consider that when using web-based services for predictions, disclosure of proprietary information is not recommended as most services do not guarantee the confidentiality of the data uploaded to their servers.

Challenges and opportunities

Data is emerging as one of the key commodities of modern drug discovery. This poses a challenge to small institutes, which normally do not have large amounts of in-house data. Nevertheless, the first step in a data strategy is to leverage whatever data is available in-house. It is therefore important to set-up rigorous ways of storing the data that is generated in a format that is searchable and suitable for subsequent analysis. Commercial data management systems like Dotmatics and Collaborative Drug Discovery Vault are efficient ways to capture the range of data generated from drug discovery projects.[79] These systems also future-proof the organization, preventing loss of data when a member of staff moves on. Data management systems are significantly more expensive than computational modeling tools. However, they offer excellent return on investment by preventing data loss, minimizing time spent finding data, and maximizing the amount of information that can be extracted from the data.

Even when leveraging all of the data generated in-house, most smaller institutions will find that there is an overall lack of data to base modeling on. Key to mitigate this is the plethora of publicly available databases. A selection of useful databases, many of which are discussed in the previous sections, are presented in Table 1.

Selecting the appropriate software is another key task requiring careful consideration. The reality is that software can represent a significant cost while a large software collection also adds complexity. Any purchase should therefore fill a specific function. Some vendors provide all

functionality in one package while others sell individual modules, so verify that any package includes the functionality that you need before purchasing. In our experience, using one commercial suite as the core and supplementing this with key bits of free software is an affordable but powerful setup. It is possible to set up an entire discovery pipeline using only free software, but this comes at the cost of complexity and sometimes performance. A list of commonly used software is provided in Table 2.

There are also many web tools available to carry out a range of computational chemistry tasks, for example, pKa predictors, P450 metabolism site predictors, etc. An important issue to consider when employing these tools is whether your data and IP are secure. Many require uploading structures on a website, and this may allow others to see your compounds.

In addition to the software, some hardware is required. However, most tasks can be accomplished using standard hardware. A good setup is a high-end workstation with a good graphics card coupled with a simple server for licenses and hosting web applications. For a workstation, the choice of operating system (OS) is a matter of personal preference. Choice of software may dictate the OS required, but all software packages listed in Table 2 run on Windows and Linux. In our experience, for servers Linux is generally better.

A big challenge for computational chemists in small settings is the wide range of skills required. Whereas larger set-ups may have separate bioinformatics, chemoinformatics, modeling, and IT specialists, in small settings one person may have to cover all these disciplines. Luckily, there are now many training resources available online, which can help gain the skills required. Many software vendors, including CCG, Cresset, Optibrium, and Schrödinger organize free or low-cost webinars, seminars, and user group meetings to train their users. They post many of the lectures given at these events on their websites or YouTube. Short videos of presentations showing how to tackle specific tasks and tutorials to work through are also available from most software providers' websites (Table 2) and for most databases in Table 1. RDKit, Knime, and many other open tools have very active user communities who help each other through forums, and share tools they have developed**.** Time spent on these training resources is worthwhile in our experience, as it helps with choosing the correct

settings and understanding the limitations of tools. Important learnings can also be had from networking with the wider modeling community.

Conclusions and outlooks

Computational methods play a role in the entire drug discovery and development cascade, from finding the right targets to statistical analysis of clinical data. Although smaller actors in the drug discovery area may struggle to implement all state-of-the-art techniques, key aspects can be covered using only modest resources. Throughout the perspective, we have described what we believe are the best practices for these methods as well as how they fit in the drug discovery cascade. Additionally, it is our experience that when computational scientists are closely integrated into the day-to-day activities they can influence the culture around and uptake of computational methods in the drug discovery process, and may so mitigate some of the challenges imposed by limited resources.

It is our experience that having a computational scientist on-board not only enables the various computational drug discovery approaches discussed in this Perspective but also has the potential to deliver more unforeseen benefits such as a more robust and streamlined data handling across the organization (is your team still routinely spending hours making calculations in Excel spreadsheets?) and an increased ability to leverage public data.

In conclusion, we anticipate that computational methods will play an increasingly important role in modern drug discovery both in pharma settings and across smaller institutes. Approaches that leverage the most value from computational techniques and from both internal and public data will be a key determinant of the success for many academic groups and small biotech companies.

Acknowledgments

We would like to thank Dr John Skidmore, Dr Steve Andrews, Prof Paul Whiting, Prof Paul Fish, Dr John Davis, and Prof Paul Brennan for helpful comments on the manuscript.

The Alzheimer's Research UK Drug Discovery Alliance is funded by Alzheimer's Research UK (registered charity No. 1077089 and SC042474).

Biography

Henriëtte Willems is a Senior Research Associate at the ALBORADA Drug Discovery Institute at the University of Cambridge. She obtained her Ph.D. in 1996 from the University of Cambridge, after which she worked for several small biotechs and large pharma companies, including Astex Pharmaceuticals, GSK, and Takeda. She has published on several structure-based drug design projects and contributed to drug discovery patents in a wide range of therapeutic areas.

Stephane De Cesco is a Senior Research Associate at the Alzheimer's Research UK Oxford Drug Discovery Institute. He was awarded his Ph.D. in 2016 from McGill University, Montreal, Canada on the topic of combining computational, synthetic and biophysical tools for the discovery of reversible covalent inhibitors.

Fredrik Svensson is a Senior Research Associate at the Alzheimer's Research UK UCL Drug Discovery Institute. He was awarded his Ph.D. in 2015 from Uppsala University, Sweden, after which he obtained a postdoctoral fellowship from the Swedish Pharmaceutical Society for research in cheminformatics at the University of Cambridge with Dr Andreas Bender. He has published extensively on the use of machine learning in drug discovery.

Corresponding Author Information: f.svensson@ucl.ac.uk

Abbreviations Used:

VS:Virtual Screening; MMP: Matched Molecular Pairs; ML: Machine Learning; FEP: Free Energy Pertubation; MM-GBSA:molecular mechanics generalized Born surface area; MM-PBSA:molecular mechanics Poisson–Boltzmann surface area; LLE: Lipofilic Ligand Efficiancy; QM: quantum-mechanical

References

(1) Galkina Cleary, E.; Beierlein, J. M.; Khanuja, N. S.; McNamee, L. M.; Ledley, F. D. Contribution of NIH Funding to New Drug Approvals 2010–2016. *Proc. Natl. Acad. Sci.* **2018**, *115*, 2329–2334.

(2) Frearson, J.; Wyatt, P. Drug Discovery in Academia: The Third Way? *Expert Opin. Drug Discov.* **2010**, *5*, 909–919.

(3) Tralau-Stewart, C.; Low, C. M. R.; Marlin, N. UK Academic Drug Discovery. *Nat. Rev. Drug Discov.* **2013**, *13*, 15.

(4) Frye, S.; Crosby, M.; Edwards, T.; Juliano, R. US Academic Drug Discovery. *Nat. Rev. Drug Discov.* **2011**, *10*, 409.

(5) Shanks, E.; Ketteler, R.; Ebner, D. Academic Drug Discovery within the United Kingdom: A Reassessment. *Nat. Rev. Drug Discov.* **2015**, *14*, 510.

(6) Morrison, C. Fresh from the Biotech Pipeline—2018. *Nat. Biotechnol.* **2019**, *37*, 118–123.

(7) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discov.* **2013**, *12*, 948–962.

(8) Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M. The Multiple Roles of Computational Chemistry in Fragment-Based Drug Design. *J. Comput. Aided. Mol. Des.* **2009**, *23*, 459–473.

(9) Katsila, T.; Spyroulias, G. A.; Patrinos, G. P.; Matsoukas, M.-T. Computational Approaches in Target Identification and Drug Discovery. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 177–184.

(10) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.

(11) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* **2018**, *20*, 58.

(12) Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful Applications of Computer Aided Drug

Discovery: Moving Drugs from Concept to the Clinic. *Curr. Top. Med. Chem.* **2010**, *10*, 127–141.

(13) Behan, F. M.; Iorio, F.; Picco, G.; Gonçalves, E.; Beaver, C. M.; Migliardi, G.; Santos, R.; Rao, Y.; Sassi, F.; Pinnelli, M.; Ansari, R.; Harper, S.; Jackson, D. A.; McRae, R.; Pooley, R.; Wilkinson, P.; van der Meer, D.; Dow, D.; Buser-Doepner, C.; Bertotti, A.; Trusolino, L.; Stronach, E. A.; Saez-Rodriguez, J.; Yusa, K.; Garnett, M. J. Prioritization of Cancer Therapeutic Targets Using CRISPR–Cas9 Screens. *Nature* **2019**, *568*, 511–516.

(14) Koscielny, G.; An, P.; Carvalho-Silva, D.; Cham, J. A.; Fumis, L.; Gasparyan, R.; Hasan, S.; Karamanis, N.; Maguire, M.; Papa, E.; Pierleoni, A.; Pignatelli, M.; Platt, T.; Rowland, F.; Wankar, P.; Bento, A. P.; Burdett, T.; Fabregat, A.; Forbes, S.; Gaulton, A.; Gonzalez, C. Y.; Hermjakob, H.; Hersey, A.; Jupe, S.; Kafkas, Ş.; Keays, M.; Leroy, C.; Lopez, F.-J.; Magarinos, M. P.; Malone, J.; McEntyre, J.; Munoz-Pomer Fuentes, A.; O'Donovan, C.; Papatheodorou, I.; Parkinson, H.; Palka, B.; Paschall, J.; Petryszak, R.; Pratanwanich, N.; Sarntivijal, S.; Saunders, G.; Sidiropoulos, K.; Smith, T.; Sondka, Z.; Stegle, O.; Tang, Y. A.; Turner, E.; Vaughan, B.; Vrousgou, O.; Watkins, X.; Martin, M.-J.; Sanseau, P.; Vamathevan, J.; Birney, E.; Barrett, J.; Dunham, I. Open Targets: A Platform for Therapeutic Target Identification and Validation. *Nucleic Acids Res.* **2017**, *45*, D985–D994.

(15) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2016**, *45*, D158–D169.

(16) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(17) Gilbert, J.; Pearcy, N.; Norman, R.; Millat, T.; Winzer, K.; King, J.; Hodgman, C.; Minton, N.; Twycross, J. Gsmodutils: A Python Based Framework for Test-Driven Genome Scale Metabolic

Model Development. *Bioinformatics* **2019**, *35*, 3397–3403.

(18) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(19) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. a; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10*, 188–195.

(20) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. *ACS Chem. Biol.* **2016**, *11*, 1255–1264.

(21) Meanwell, N. A. Improving Drug Design: An Update on Recent Applications of Efficiency Metrics, Strategies for Replacing Problematic Elements, and Compounds in Nontraditional Drug Space. *Chem. Res. Toxicol.* **2016**, *29*, 564–616.

(22) Gilad, Y.; Nadassy, K.; Senderowitz, H. A Reliable Computational Workflow for the Selection of Optimal Screening Libraries. *J. Cheminform.* **2015**, *7*, 61.

(23) Paricharak, S.; Méndez-Lucio, O.; Ravindranath, A. C.; Bender, A.; IJzerman, A. P.; van Westen, G. J. P. Data-Driven Approaches Used for Compound Library Design, Hit Trage and Bioactivity Modeling in High-Throughput Screening. *Brief. Bioinform.* **2018**, *19*, 277–285.

(24) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435–444.

(25) Vidler, L. R.; Watson, I. A.; Margolis, B. J.; Cummins, D. J.; Brunavs, M. Investigating the Behavior of Published PAINS Alerts Using a Pharmaceutical Company Data Set. *ACS Med.*

Chem. Lett. **2018**, *9*, 792–796.

(26) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36–44.

(27) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341.

(28) ChemAxon Suite. ChemAxon 2019.

(29) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery. *ACS Chem. Neurosci.* **2016**, *7*, 767–775.

(30) Zhu, T.; Cao, S.; Su, P.-C.; Patel, R.; Shah, D.; Chokshi, H. B.; Szukala, R.; Johnson, M. E.; Hevener, K. E. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis. *J. Med. Chem.* **2013**, *56*, 6560–6572.

(31) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discov. Today* **2004**, *9*, 430–431.

(32) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided. Drug Des.* **2011**, *7*, 146–157.

(33) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.

(34) Pierce, A. C.; Jacobs, M.; Stuver-Moody, C. Docking Study Yields Four Novel Inhibitors of the Protooncogene Pim-1 Kinase. *J. Med. Chem.* **2008**, *51*, 1972–1975.

(35) Damm-Ganamet, K. L.; Arora, N.; Becart, S.; Edwards, J. P.; Lebsack, A. D.; McAllister, H. M.; Nelen, M. I.; Rao, N. L.; Westover, L.; Wiener, J. J. M.; Mirzadegan, T. Accelerating Lead

Identification by High Throughput Virtual Screening: Prospective Case Studies from the Pharmaceutical Industry. *J. Chem. Inf. Model.* **2019**, *59*, 2046–2062.

(36) Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. Molecular Docking Screens Using Comparative Models of Proteins. *J. Chem. Inf. Model.* **2009**, *49*, 2512–2527.

(37) Forli, S. Charting a Path to Success in Virtual Screening. *Molecules* **2015**, *20*, 18732–18758.

(38) Broccatelli, F.; Brown, N. Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2014**, *54*, 1634–1641.

(39) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895–2907.

(40) Davis, A. M.; St-Gallay, S. A.; Kleywegt, G. J. Limitations and Lessons in the Use of X-Ray Structural Information in Drug Design. *Drug Discov. Today* **2008**, *13*, 831–841.

(41) Skuta, C.; Popr, M.; Muller, T.; Jindrich, J.; Kahle, M.; Sedlak, D.; Svozil, D.; Bartunek, P. Probes & Drugs Portal: An Interactive, Open Data Resource for Chemical Biology. *Nat. Methods* **2017**, *14*, 759–760.

(42) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213.

(43) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400–D412.

(44) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2015**, *44*, D1220–

D1228.

(45) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867–881.

(46) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.

(47) Kuhn, B.; Guba, W.; Hert, J.; Banner, D.; Bissantz, C.; Ceccarelli, S.; Haap, W.; Körner, M.; Kuglstatter, A.; Lerner, C.; Mattei, P.; Neidhart, W.; Pinard, E.; Rudolph, M. G.; Schulz-Gasch, T.; Woltering, T.; Stahl, M. A Real-World Perspective on Molecular Design. *J. Med. Chem.* **2016**, *59*, 4087–4102.

(48) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(49) Bender, A.; Bojanic, D.; Davies, J. W.; Crisman, T. J.; Mikhailov, D.; Scheiber, J.; Jenkins, J. L.; Deng, Z.; Hill, W. A. G.; Popov, M.; Jacoby, E.; Glick, M. Which Aspects of HTS Are Empirically Correlated with Downstream Success? *Current Opinion in Drug Discovery and Development*. **2008**, 11, 327-37.

(50) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53*, 4891–4905.

(51) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 221–234.

(52) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided. Mol. Des.* **2018**, *32*, 1–20.

(53) Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **2016**, *56*, 1063–1077.

(54) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895–913.

(55) Yuriev, E.; Holien, J.; Ramsland, P. A. Improvements, Trends, and New Ideas in Molecular Docking: 2012–2013 in Review. *J. Mol. Recognit.* **2015**, *28*, 581–604.

(56) Ericksen, S. S.; Wu, H.; Zhang, H.; Michael, L. A.; Newton, M. A.; Hoffmann, F. M.; Wildman, S. A. Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2017**, *57*, 1579–1590.

(57) Svensson, F.; Karlén, A.; Sköld, C. Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. *J. Chem. Inf. Model.* **2012**, *52*, 225–232.

(58) Muthas, D.; Sabnis, Y. A.; Lundborg, M.; Karlén, A. Is It Possible to Increase Hit Rates in Structure-Based Virtual Screening by Pharmacophore Filtering? An Investigation of the Advantages and Pitfalls of Post-Filtering. *J. Mol. Graph. Model.* **2008**, *26*, 1237–1251.

(59) Hawkins, P. C. D.; Stahl, G. Ligand-Based Methods in GPCR Computer-Aided Drug Design. In *Computational Methods for GPCR Drug Discovery*; Heifetz, A., Ed.; Springer New York: New York, NY, 2018; pp 365–374.

(60) Perola, E. Minimizing False Positives in Kinase Virtual Screens. *Proteins Struct. Funct.*

Bioinforma. **2006**, *64*, 422–435.

(61) Atkinson, B. N.; Steadman, D.; Mahy, W.; Zhao, Y.; Sipthorp, J.; Bayle, E. D.; Svensson, F.; Papageorgiou, G.; Jeganathan, F.; Frew, S.; Monaghan, A.; Bictash, M.; Yvonne Jones, E.; Fish, P. V. Scaffold-Hopping Identifies Furano[2,3-d]Pyrimidine Amides as Potent Notum Inhibitors. *Bioorg. Med. Chem. Lett.* **2019**, 126751.

(62) Pu, C.; Yan, G.; Shi, J.; Li, R. Assessing the Performance of Docking Scoring Function, FEP, MM-GBSA, and QM/MM-GBSA Approaches on a Series of PLK1 Inhibitors. *Medchemcomm* **2017**, *8*, 1452–1458.

(63) Verlinde, C. L. M. J.; Hol, W. G. J. Structure-Based Drug Design: Progress, Results and Challenges. *Structure* **1994**, *2*, 577–587.

(64) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(65) Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*.

(66) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(67) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488.

(68) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

(69) Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. *J.*

Chem. Inf. Model. **2019**, *59*, 4150–4158.

(70) Lombardo, F.; Desai, P. V; Arimoto, R.; Desino, K. E.; Fischer, H.; Keefer, C. E.; Petersson, C.; Winiwarter, S.; Broccatelli, F. In Silico Absorption, Distribution, Metabolism, Excretion, and Pharmacokinetics (ADME-PK): Utility and Best Practices. An Industry Perspective from the International Consortium for Innovation through Quality in Pharmaceutical Development. *J. Med. Chem.* **2017**, *60*, 9097–9113.

(71) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Mannhold, R., Kubinyi, H., Folkers, G., Oprea, T. I., Eds.; Methods and Principles in Medicinal Chemistry; 2005.

(72) Tyrchan, C.; Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 86–90.

(73) Heightman, T. D.; Callahan, J. F.; Chiarparin, E.; Coyle, J. E.; Griffiths-Jones, C.; Lakdawala, A. S.; McMenamin, R.; Mortenson, P. N.; Norton, D.; Peakman, T. M.; Rich, S. J.; Richardson, C.; Rumsey, W. L.; Sanchez, Y.; Saxty, G.; Willems, H. M. G.; Wolfe, L.; Woolford, A. J.-A.; Wu, Z.; Yan, H.; Kerns, J. K.; Davies, T. G. Structure–Activity and Structure–Conformation Relationships of Aryl Propionic Acid Inhibitors of the Kelch-like ECH-Associated Protein 1/Nuclear Factor Erythroid 2-Related Factor 2 (KEAP1/NRF2) Protein–Protein Interaction. *J. Med. Chem.* **2019**, *62*, 4683–4702.

(74) Neese, F. The ORCA Program System. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(75) Berteotti, A.; Vacondio, F.; Lodola, A.; Bassi, M.; Silva, C.; Mor, M.; Cavalli, A. Predicting the Reactivity of Nitrile-Carrying Compounds with Cysteine: A Combined Computational and Experimental Study. *ACS Med. Chem. Lett.* **2014**, *5*, 501–505.

(76) Czodrowski, P. HERG Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240–2251.

(77) Golbamaki, A.; Benfenati, E. In Silico Methods for Carcinogenicity Assessment BT - In Silico

Methods for Predicting Drug Toxicity; Benfenati, E., Ed.; Springer New York: New York, NY, 2016; pp 107–119.

(78) Kazmi, S. R.; Jun, R.; Yu, M.-S.; Jung, C.; Na, D. In Silico Approaches and Tools for the Prediction of Drug Metabolism and Fate: A Review. *Comput. Biol. Med.* **2019**, *106*, 54–64.

(79) Hohman, M.; Gregory, K.; Chibale, K.; Smith, P. J.; Ekins, S.; Bunin, B. Novel Web-Based Tools Combining Chemistry Informatics, Biology and Social Networks for Drug Discovery. *Drug Discov. Today* **2009**, *14*, 261–270.

Table of Contents Graphics

Table 1. Databases to mine for compound activity data and/or target–disease links.

Type	Name	Molecule search	Target search	Disease search	Description
Internet search by molecule	chemspider.com	y	n	n	Converts names and IDs to structures; links to vendors and bioactivity data
	isciencesearch.com	y	n	n	Searches the internet by (sub)structure or name for supplier; synthesis; bioactivity; literature; patent or toxicity data
Large databases of chemical matter with published activity data	bindingdb.org	y	y	y	Published binding affinities for drug-like molecules
	ebi.ac.uk/chembl	y	y	n	Assay data on chemical matter extracted from literature
	drugtargetcommons.fimm.fi	y	y	n	Published binding affinities for drug-like molecules; contains some data not in ChEMBL
	solr.ideaconsult.net/search/excape	y	y	n	ChEMBL and PubChem combined
	pubchem.ncbi.nlm.nih.gov	y	y	n	Assay data from literature and NIH screens
	cansarblack.icr.ac.uk	y	y	y	Chemical and pharmacological data for over one million bioactive small molecules focused on cancer targets
Annotated tool	probes-drugs.org	y	y	n	Lots of probes and drugs with information on selectivity; potency and places to buy

compound databases	Guidetopharmacology.org	y	y	y	Probe and target information
	chemicalprobes.org	y	y	n	Chemical probes with reviews and ratings
	probeminer.icr.ac.uk	n	y	n	Compares chemical probes for a target
Drug databases	cheminfo.charite.de/superdrug2	y	n	n	Knowledge-base of approved and marketed drugs
	drugcentral.org	y	y	y	Drug compendium integrating structure; bioactivity; regulatory; pharmacologic actions and indications
	db.idrblab.net/ttd	y	y	y	~5;000 Patented drugs and their targets; disease area and phase information
Miscellaneous	surechembl.org	y	y	y	Patent database with text and (sub)structure search facility
	ebi.ac.uk/pdbe	y	y	n	Crystal; CryoEM; and NMR structures of proteins
	uniprot.org	n	y	n	Database of proteins; sequences and domains; useful for homology searches
	opentargets.org	n	y	y	Links targets to diseases and vice versa; scoring based on evidence for link

Table 2. Commonly used software for key computational drug discovery tasks.

Maker	Name	Cost	Compound filtering	2D to 3D	Pharmacophore or shape search	Docking	Visualization	Clustering	Chemical Database	MMP	Conformer analysis	QM
Scripps	AutoDock	Free				y						
BioSolveIT		Charge			InfiniSee/FlexS/FTrees	FlexX	2D/3D					
CCDC	CSD-Discovery	Charge		y	y	Gold	2D/3D				y	
CCG	MOE	charge	y	y	y	y	3D	y		y	y	can link
ChemAxon		Free Ac*	cxcalc	molconvert			2D/3D; Marvin	JKlustor	JChem		Marvin	
Cresset	Blaze	Charge			Blaze		2D/3D					
Dotmatics	Browser;	Charge	Vortex				2D	Vortex	Browser	Vortex		

	Studies											
Knime	Knime Analytics	Free	y	y			2D	y	can link	y	y	
OpenEye	OEsuite	Charge	FILTER	QUACPAC/OMEGA	ROCS	OEdocking	2D/3D; VIDA				FREEFORM	
Optibrium	StarDrop	Charge	y				2D	y	can link	y		
OSIRIS	Data Warrior	Free	y				2D	y	can link			
RDKit	RDKit	Free	y	y			y	y	can link		y	
Schröding er	Maestro suite	Charge	y	LigPrep	Phase	Glide	2D/3D	y (Canvas)	LiveDesign		MacroModel	Jaguar

*These vendors offer the licenses free of charge to academic groups (subject to contract).