Maximizing the integration of virtual and experimental screening in hit discovery

Dávid Bajusz, György Miklós Keserű*

Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Magyar tudósok krt 2, H-1117 Budapest

*Corresponding author.

Dávid Bajusz, PhD

research fellow

ORCID: https://orcid.org/0000-0003-4277-9481

*György Miklós Keserű, DSc, corr. member of Hung. Acad. Sci.

research professor

ORCID: https://orcid.org/0000-0003-1039-7809

Email: keseru.gyorgy@ttk.hu

Phone: +36 1 3826 821

Structured abstract

Introduction: Experimental and virtual screening contributes to the discovery of more than 50% of clinical candidates. Considering the similar concept and goals, early phase drug discovery would benefit from the effective integration of these approaches.
Areas covered: After reviewing the recent trends in both experimental and virtual screening, we discuss different integration strategies from parallel, focused, sequential and iterative screening. Strategic considerations are demonstrated in a number of real-life case studies.
Expert opinion: Experimental and virtual screening are complementary approaches that should be integrated in lead discovery settings. Virtual screening can access extremely large synthetically feasible chemical space that can be effectively searched on GPU clusters or cloud architectures. Experimental screening provides reliable datasets by quantitative HTS applications, and DNA-encoded libraries (DEL) have enlarged the chemical space covered by these technologies. These developments, together with the use of artificial intelligence methods, represent new options for their efficient integration. The case studies discussed here demonstrate the benefits of complementary strategies such as focused and iterative screening.

Keywords: artificial intelligence, DNA-encoded libraries, focused screening, high throughput screening, iterative screening, parallel screening, phenotypic screening, sequential screening, ultra-large screening, virtual screening

Article highlights

- Large virtual databases of synthetically accessible compounds cover the larger part of the chemical space available for virtual screening
- DNA-encoded libraries (DELs) represent better coverage of the chemical space for experimental screening and provide large training datasets for virtual screening
- Quantitative high-throughput screening generates better quality data available for the iterative improvement of virtual screening protocols
- GPU clusters and cloud architectures, together with parallelized software applications and artificial intelligence techniques, provide higher performance for virtual screening
- Iterative integration of experimental and virtual screening maximizes the benefits of these complementary approaches

1. Introduction

Lead generation is in the heart of the preclinical drug discovery process, since the main structural features of the lead molecules were shown to be kept during the optimization leading to clinical candidates [1]. A recent analysis of the clinical candidates revealed that 43% of small molecule drug candidates come from already existing compounds (already synthesized and characterized compounds, but not necessarily drugs) [2]. Although this approach would speed up discovery programs and is well suited for identifying fast followers, the limited freedom to operate would make navigation in the IP space difficult and risky. Screening technologies, however, provide new chemical matter that are outside these limitations. In fact, the other 57% of the leads originated from screening approaches including high-throughput screening (HTS, 29%), virtual screening (VS, 14%), focused screens (8%), fragment screens (5%) and screening of DNA encoded libraries (1%). These figures suggest that in addition to experimental screening technologies, VS can contribute significantly to hit and lead discovery. Being the second after HTS, VS approaches have identical goals and objectives and therefore should not

be treated as competitive, but highly complementary to experimental screening [3]. Consequently, early stage drug discovery programs could benefit from their integration. Focused screening, the third largest screening contributor of drug leads, reinforces the need for the effective integration. There are multiple options to integrate these technologies. In the order of integration level, these include the parallel approach, focused, sequential and iterative screening. Parallel screening uses experimental and virtual screening protocols on the same compound library. While HTS typically operates on large diverse screening decks, virtual screening can be used to focus the collection to a preselected subset of compounds, or even compounds not contained in the screening collection. Consequently, focused screening would provide higher hit rates realized with a diverse set of chemotypes. The next level of integration is sequential screening that attempts to maximize the cost-benefit ratio of the campaign by balancing the two technologies. Finally, the most integrated approach is iterative screening that utilizes the SAR knowledge generated by focused or sequential screening to redesign the screening library for maximizing the information content of the hits.

We have discussed the integration strategies of VS and HTS ten years ago and collected a number of case studies for their effective integration in different scenarios [4]. Another contemporary review was also published on this topic, with more emphasis on virtual screening [5]. In the past ten years, the field has undergone major paradigm shifts. This review will therefore concentrate on introducing the rapidly changing landscape of VS and HTS, with recent case studies to illustrate their synergy in today's AI-dominated applications.

2. Virtual screening

Parallel to the rapid increase of computational capacity, virtual screening has originally appeared as an inexpensive alternative to high-throughput screening, unlocking access to a much wider (virtual) chemical space than what was physically available. The rationalization of drug development costs, as well as the democratization of early drug discovery is still the main driving force in developing and applying computational methods in the hit discovery, hit-to-lead and lead optimization processes [6]. Traditionally, VS methods are grouped into ligand-based and (protein) structure-based methods, depending on whether the structure of the therapeutic (protein) target is utilized during the screening process. VS methods can be combined into intricate, step-wise or parallel screening workflows. In fact, some ligand-based methods are primarily utilized as fast pre-filtering steps prior to the main screening step that usually aims to prioritize/select compounds with on-target activity. These methods can be easily combined with experimental screening efforts as well, for more focused experimental screens or smarter library design that does not necessarily require HTS infrastructure. Since a detailed listing of virtual screening software is outside the scope of this review, we point to a recently published work with an exhaustive collection of online virtual screening tools [7]. In contrast to stand-alone software suites, these are less customizable webservers, representing a good starting point for non-expert practitioners.

Large-scale filtering of physically available or synthetically accessible compound databases is most effectively realized using molecular descriptors. Molecular descriptors are numerical values that encode various features/characteristics of compounds, ranging from core physico-chemical properties like molecular weight or lipophilicity (usually expressed as logP, where P is the octanol-water partition coefficient) to more intricate topological descriptors like the Randic index [8]. They can be efficiently calculated even for large databases, and therefore they are ideal choices to define filtering criteria to cut down on dataset size prior to the more computationally expensive screening steps. Since the well-known "rule of five" by Lipinski et

al. [9], many descriptor-based filtering rules have been implemented to produce ligand sets focused by molecular size (chiefly, drug-like, lead-like and fragment filters, corresponding to various stages of early drug discovery), with some prominent examples collected in Table 1. It is worth to note that most of these rules utilize the molecular weight as a proxy for compound size, but recent trends have realized the superiority of the number of heavy (non-hydrogen) atoms for this purpose. For example, in our recent work [10] we apply only this criteria to capture the fragment size range defined by Murray and Rees [11]. Another important group of filtering rules aims to assess compound "quality", *i.e.* the reliability of certain compound classes when they show up as screening hits (Table 1). In contrast to the descriptor-based REOS filters, the PAINS (Pan-Assay Interference Compounds) rules of Baell and Holloway explicitly define certain substructure classes that frequently show up as false positives in screening campaigns [12]. These substructure filters have been canonicalized as a "must" in modern VS efforts: most current molecular modeling/cheminformatics software offer a convenient way to remove PAINS from screening sets, and certain journals explicitly require this [13]. The issue of PAINS was even highlighted in a witfully illustrated comment in Nature [14]. In addition to the mentioned, more wide-spread filtering rules, descriptor-based scoring schemes/multi-criteria optimization methods can be used to compile focused libraries against prominent target classes, such as GPCRs [15] or kinases [16]. Since multi-criteria optimization is a very popular approach in drug discovery in general, we can point the reader to excellent reviews on the topic [17–19].

Table 1. Some of the most well-known filtering rules for virtual screening/library design

	Rules[1,2]	Reference
Size-based rules		
Drug-like	150 <= MW <= 500 logP <= 5 rotB <=7 PSA < 150 Hb_donors <= 5 Hb_acceptors <= 10	Lipinski et al. [9] "Rule of 5"
	160 <= MW <= 480 -0.4 <= logP <= 5.6 20 <= N_atoms <= 70 40 <= MR <= 130	Ghose et al. [20]
	$QED = e^{\frac{1}{n}\sum_{i=1}^{n} \ln d_i}$	Bickerton et al. [21] "Quantitative estimate of drug-likeness" (QED)
Lead-like	250 <= MW <= 350 logP <= 3.5 rotB <= 7	Teague et al. [22]
Fragment	MW <= 250 logP <= 3.5 rotB <=5	Carr et al. [23]
	MW <= 300 logP <= 3 Hb_donors <= 3 Hb_acceptors <= 3	Congreve et al. [24] "Rule of 3"

	rotB <=3 *PSA <= 60*	
	10 <= N_heavy <= 16 *140 <= MW <= 230* *0 <= logP <= 2*	Murray and Rees [11]
Quality-based rules		
REOS	200 <= MW <= 500 -5 <= logP <= 5 Hb_donors <= 5 Hb_acceptors <= 10 -2 <= f. charge <= 2 rotB <= 8 15 <= N_heavy <= 50	Walters and Namchuk [25] "Rapid Elimination of Swill"
PAINS	SMARTS patterns for substructure filter families A (16), B (55) and C (409)	Baell and Holloway [12] "Pan-Assay Interference Compounds"

[1] MW: molecular weight, logP: (calculated) logarithm of the octanol-water partition coefficient, rotB: number of rotatable bonds, PSA: polar surface area, Hb_donors/acceptors: number of hydrogen-bond donor/acceptor groups, N_atoms: number of atoms, MR: molar refractivity, d_i: desirability functions based on eight pre-defined descriptors, N_heavy: number of heavy (non-hydrogen) atoms, f.charge: formal charge

[2] Optional criteria are marked with italics.

In the last ten years, the driving force behind the development of VS methodology was the possibility of accessing larger and larger chemical spaces. While the chemical space of possible "druglike" molecules (MW <= 500) is estimated to be in the order of 10^{60} compounds [26, 27], "traditional" approaches enabled access to 10^5-10^6 compounds in big pharma for HTS, and 10^6-10^7 compounds for the virtual screening of physically available, in-stock compounds. The research community (compound vendors, big pharma, as well as academic groups) has realized that it is possible to tap into chemical spaces of 10^8-10^{10} compounds by systematically generating virtual compound databases through the combination of existing chemical building blocks with robust chemical reactions [28]. Such databases of synthetically easily accessible compounds include vendor databases like Enamine's REAL (readily accessible) database [29] and aggregators such as Mcule's ULTIMATE [30], or the latest version of the popular ZINC database [31]. This approach was taken even further in big pharma, starting with the reaction-driven chemical space generated at Pfizer [32], with recent proprietary virtual libraries reaching up to 10^{20} compounds [33]. It is worth to note that even larger chemical spaces are conceivable based on a purely rule-based ("whatever is correct on paper") approach, without regard to synthetic accessibility: the latest iteration of the Reymond group's generated database (GDB-17) contains 166 billion ($1.66 * 10^{11}$) virtual compounds with an upper limit of 17 heavy atoms [34]. Figure 1 summarizes the general hierarchy of chemical spaces available for virtual screening.

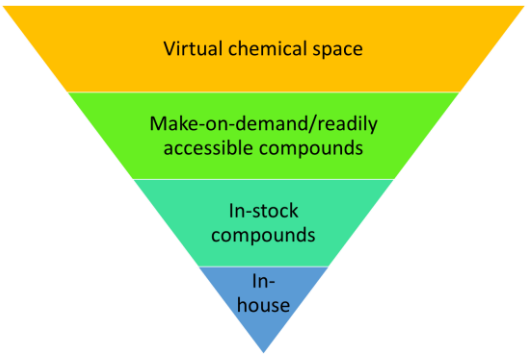

Figure 1. General hierarchy of chemical spaces available for VS. In-house datasets typically include 10^3-10^4 compounds for academic research groups and 10^5-10^6 compounds for big pharma. In-stock commercially available compounds typically list 10^6-10^7 compounds on major aggregator sites. The scope of on-demand commercial compounds is quickly growing; currently it is in the magnitude of 10^8-10^{10} compounds. Virtual chemical space is practically unlimited and can only be quantified with certain constraints (*e.g.* 10^{60} compounds with MW <= 500, or 1.66 * 10^{11} compounds with max. 17 heavy atoms).

In the meantime, the computing power has increased steadily: high-performance computing (HPC) infrastructure is now widely available from commercial providers such as Amazon AWS or Google Cloud, and through funding agencies, as in the case of PRACE [35]. To keep up with the heavier workloads of ever larger virtual screens, the popular docking software Autodock 4 was extended to GPU infrastructure, providing speedups of multiple orders of magnitude in comparison to the traditional CPU version [36]. In fact, a recent application has showcased the possibility to dock one billion compounds in a single day with Autodock-GPU, on HPC infrastructure [37]. Guidelines for large-scale docking based on the DOCK software and the ZINC database were recently published as well [38]. Similar speedups are observed for FastROCS [39], the GPU-based version of the popular shape similarity screening software ROCS (Rapid Overlay of Chemical Structures) [40, 41], and computationally efficient algorithms for pharmacophore screening were also reported [42], and even integrated with the ZINC database [43]. These advances resulted in the possibility to screen libraries in the magnitude of 100 million compounds or even more: the term “ultra-large virtual screening” was coined to distinguish these efforts from the earlier generations of VS methodologies. A recent opinion piece by Derek Lowe nicely illustrates the general progress of this field as a delicate balance between the increasing computational demand and new, more efficient algorithmic ideas [44]. The recent open-source software VirtualFlow provides a flexible platform for ultra-large virtual screens with support for the most popular file formats and multiple docking algorithms [45]. A recent review by Murugan *et al.* provides a summary of parallelized docking software, suitable for high-performance computing – along with an introduction to different software parallelization schemes and an outlook towards the future of quantum computing in virtual screening [46]. Successful applications of ultra-large screening were reported to result in high-affinity binders of the AmpC β-lactamase (AmpC) and the D_4 dopamine receptor [47], while the synthon-based V-SYNTHES approach links the library generation and screening steps to achieve the scale of 11 billion screened compounds, yielding hit rates around 30% and nanomolar hits against cannabinoid receptors and the ROCK1 kinase [48]. Screening the leadlike subset of the ZINC15 database (235 million compounds [49]) has resulted in the identification of potent, noncovalent inhibitors of the SARS-CoV-2 main protease [50]. It is worth to note that ligand- and structure-based virtual approaches (such as docking, shape screening and pharmacophore screening) can be flexibly combined to result in stepwise [51] or parallel [52] virtual screening workflows, although to our knowledge this is

not yet widespread in ultra-large screens. We should note that consensus screening strategies can also be realized on the level of considering multiple scoring functions in docking-based virtual screening workflows [53, 54].

Another strong trend in virtual screening, as well as research in general, is the adaptation of artificial intelligence and machine learning methodologies. In structure-based virtual screening, the recent introduction of AlphaFold was a major game changer, providing accurately predicted protein structures by tackling the protein folding problem [55, 56]. In the meantime, machine learning is already heavily rooted in classification/regression problems concerning targets of general interest, such as ADMET targets [57], and deep learning methodologies are gaining popularity in ligand-based and structure-based virtual screening as well [58]. Another popular usage of AI methods in virtual screening is to cut back on the computational demand of docking by training deep learning models to predict the docking scores, thus bypassing the more demanding docking step for a large portion of the database (the "low-scoring" ligands) [59, 60]. A prominent platform, termed Deep Docking was contributed recently by the Cherkasov lab [61]. The Deep Docking workflow was demonstrated by the discovery of new SARS-CoV-2 main protease inhibitors by the virtual screening of over 40 billion compounds [62], and an instructive introduction to this methodology was published earlier this year in Nature Protocols [63]. Another recent software is MolAICal, which implements deep learning tools to tackle the 3D drug design problem through a different concept: a deep learning model is trained on experimental data from the PDBbind database [64], and applied prospectively to design 3D ligand structures inside the binding pocket, thus realizing an AI-enhanced solution for *de novo* design [65].

One frequently criticized aspect of machine learning models in general is their lack of interpretability: while "black-box" models may be successful in predicting new candidates, they offer little in terms of the rationale that could be provided to support further decision making. Additional limitations of the most mainstream machine learning concepts are sensitivity to their training input (*i.e.* usually they will be reliable to predict one specific bioassay, but not "mixed" results of the same end-point determined with different assay conditions), and the lack of multi-target prediction within the same model. To deal with these limitations, an emerging trend in recent years has been the application of perturbation theory [66], combined with machine learning (PTML). Here, different assay conditions and targets can be incorporated into the model through the definition of the appropriate perturbation theory operators [67]. The method was most prominently demonstrated by the González-Díaz and Speck-Planche groups, with a wide range of applications in multi-target disease conditions like Alzheimer's disease [68], oncology [69, 70] and multi-strain antituberculosis drugs [71], and in even more complex systems, such as metabolic reaction networks [72].

Altogether, we presume that AI methods will gain even more ground in the close future, parallel to the increasing size of VS datasets.

3. Experimental screening

For decades, the development of high-throughput screening technologies was driven by miniaturization. Accessing smaller sample volumes, while maintaining accuracy, translated into the possibility of larger screens, at better cost-to-volume ratios. In fact, high-throughput screening has already reached maturity in this aspect in the recent decade, when acoustic droplet handling instruments have pushed down the sample volumes of previously high-consumption techniques, such as protein crystallography, into the nanoliter scale [73]. Together with the

previously unmatched global collaborative research efforts launched after the COVID-19 outbreak, these advances have contributed to the democratization of accessing high-end screening instrumentation by academic research groups, through initiatives like COVID Moonshot [74]. COVID Moonshot has resulted in successful X-ray fragment screens against multiple SARS-CoV-2 viral targets, with compound libraries contributed by research groups all over the world [75, 76]. Similar collaborative screening efforts have resulted in the first inhibitors of novel oncotargets, such as OTUB2 [77]. These examples fit well into the overall trend of HTS shifting from a closed and proprietary effort towards a collaborative activity within the pharma sector [78]. At the same time, somewhat counterintuitively, the number of screened compounds per HTS project seemed to decrease in the past decades [79], possibly due to the advances in smarter library design, and higher-quality HTS data. It is also noteworthy that HTS has found applications outside of drug discovery as well, such as in biotechnology, for the design of industrial microorganisms [80].

With access to nL sample volumes, the past decade has brought forth serious advancements in increasing HTS data quality. One prominent trend is the development of quantitative HTS (qHTS) methods, where data reliability is enhanced by incorporating concentration-dependent measurements as part of the primary HTS campaign [81]. Quantitative HTS has since been implemented into major, public screens such as the Tox21 initiative [82], and data analysis/fitting methods have been fine-tuned to robustly handle the quality control of qHTS results [83, 84]. In the meantime, the focus of HTS screens has shifted from mechanistic (single-target) towards phenotypic screening. Phenotypic screens typically employ cell-based endpoints to maximize the biological relevance of the resulting information [85] and therefore, they have no real alternative in virtual screening. While the mechanism of action of the resulting hits is obviously less clear, the success of phenotypic drug discovery was demonstrated by first-in-class medicines [86], exemplified by novel treatment options against spinal muscular atrophy [87], cystic fibrosis [88] and hepatitis C [89]. Methodologies were adapted to simulate tumour microenvironments in qHTS settings, as well [90]. Phenotypic screens have found successful applications in the drug discovery against infectious diseases [91], including SARS-CoV-2 [92]. Still, phenotypic HTS poses considerable challenges in hit validation and target deconvolution, which might discourage pharma companies from implementing this methodology [93, 94]. There have also been major paradigm shifts as well, including the development of human organoids: stem-cell derived 3D cell culture systems that recreate the architecture and physiology of human organs [95, 96]. At the same time, high-content screening techniques apply automated microscopy with quantitative image analysis for the high-throughput, quantitative analysis of cellular phenotypes [97].

HTS has benefited from innovative new technologies as well, such as the application of DNA-encoded libraries (DEL) [98]. First conceptualized in 1992 [99], the DEL technology has reached its maturity in the past decade, with the advent of the necessary technological advances [100]. DELs contain small molecules, which are individually coupled to unique DNA sequences, acting as “molecular barcodes”. The library is tested simultaneously in the same vessel, with the protein target immobilized and then incubated with the library members (DEL selection), which may approach billions of unique compounds, unlocking access to vast chemical spaces [101]. After different washing steps to remove non-binders, the most active binders are recovered by suitable elution procedures and their “barcodes” are read out by decoding their respective DNA tags, using high-throughput DNA sequencing after a PCR (polymerase chain reaction) amplification step (Figure 2) [102]. Significant efforts have been dedicated to perfecting the DEL technology, including the development of DNA-compatible

("on-DNA") chemical reactions [103], as well as data analysis platforms [104]. More recently, DEL selection was made available on cell surfaces, and inside of living cells [105].

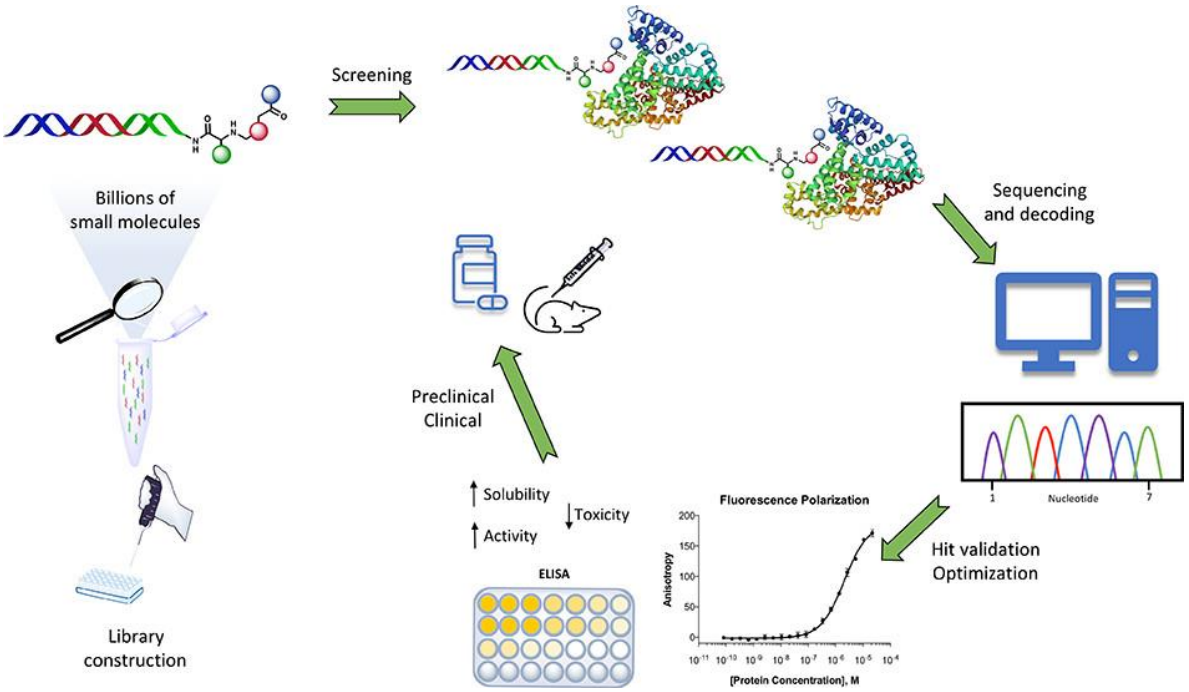

Figure 2. General drug discovery workflow employing DEL libraries. The library is constructed by coupling small molecules to unique DNA tags. During experimental screening, some ligands are bound to the target protein (hits). Hits are detected by amplifying the corresponding DNA tags with a PCR reaction, and subsequently read out by DNA sequencing. Afterwards, the hits are validated and progressed to preclinical optimization and, ideally, clinical trials. Reprinted with permission from Gironda-Martínez et al. *ACS Pharmacol. Transl. Sci.*, **2021**, *4*, 1265–1279. Copyright © **2021** American Chemical Society.

Finally, HTS was not exempt from the major scientific trend of the recent years, *i.e.* the expansion of artificial intelligence (AI). In high-throughput screening, the main entry point of AI-driven technologies is the design of smarter screening campaigns. Instead of screening the full deck of available compounds, machine learning methods can be applied to conduct iterative screening campaigns, where a smaller batch of compounds is tested and the results are fed into a machine learning algorithm that recommends structural or biological analogs of the discovered hits to be tested in the next batch [106]. By iteratively repeating these steps, the majority of the active compounds can be retrieved by screening less than half of the full compound stack in as little as three to six iterations. In another example reported by Novartis, screening approx. 1% of the full collection in ≤ 10 iterations retained diverse compounds belonging to the top 0.5% of the most active compounds [107]. With more iterations, iterative screening can identify highly active molecules even in the (relatively common) case, where the initial set contains few or no actives [108].

4. Integration of virtual and experimental screening

Various strategies exist for combining the advantages of virtual and experimental screening, although not all of them are employed to the same extent. In this section, we aim to showcase these possibilities on selected case studies of recently published hit discovery projects, mostly against prominent oncotargets.

Parallel screening

Parallel screening means the application of more screening methodologies on the same chemical library, independently of each other. Since the main goal of virtual screening is usually the rationalization of the capacity and efforts spent on experimental screening, employing a virtual and an experimental setup in parallel cannot be considered a typical concept. There are occasional examples nonetheless, mostly with the aim to provide a methodological comparison on the advantages and complementarities of these methods. One such example is the discovery of new inhibitors of the CDP-ME kinase (IspE), an antibacterial target [109]. Here, the authors have used their in-house virtual library of over 4 million compounds for a stepwise virtual screening cascade, finally purchasing and testing 14 compounds. In parallel, they have submitted a focused kinase-specific library of over 6,000 compounds for experimental screening. In addition, the HTS library was also submitted to virtual screening. Ultimately, besides identifying several new IspE inhibitors, the authors have dedicated a detailed discussion to the two methodologies, including the comparison of the retrieved hits by their chemical structures, as well as their ranks in the two hitlists. We also point to a classic example from the Shoichet lab [110], but we refrain from a more detailed summary here, as it was reviewed earlier [4].

In contrast to the above scenario, the parallel application of orthogonal virtual screening methodologies is more frequent. Here, the aim is usually to retrieve more robust results, for example by the consensus of two or more different virtual screening concepts, such as ligand docking, pharmacophore screening or shape screening. In our recent work, we have applied the three mentioned methods for the discovery of new inhibitors of the MELK kinase from our in-house compound library [52]. MELK (Maternal Embryonic Leucine-zipper Kinase) is an oncotarget with diverse functions in cellular processes, which was observed to be overexpressed in many human cancers [111]. Our efforts have resulted in the discovery of several new MELK inhibitors, with six of the primary hits sharing the same 1,2,4-triazolo[1,5-b]isoquinoline scaffold. Further exploration of the structure-activity relationship (SAR) of this series finally led to the identification of a submicromolar hit [52]. From the nine primary hits, five have resulted from a consensus of all three modeling concepts, while four of them were nominated by various combinations of two screening concepts. While the application of consensus screening is inexpensive for small libraries, it requires more thorough consideration for ultra-large screening. Nonetheless, with access to GPU workstations and HPC infrastructure, this is now feasible [37]. In fact, an early example was already published at the end of last year, where a Deep Docking workflow was implemented to screen 40 billion compounds with a consensus of five mainstream docking algorithms, resulting in the discovery of new SARS-CoV-2 main protease inhibitors [62].

In addition to multiple screening concepts and multiple docking programs, parallel virtual screening can be realized on the level of multiple scoring functions as well (consensus scoring), as exemplified by the work of Liao et al., which resulted in the first-in-class STAT5 inhibitor, IST5-002 [112]. Signal transducers and activators of transcription (STATs) are a family of seven transcription factors with key roles in intracellular signaling, recently pursued as promising oncotargets in diverse indications, such as leukemia [113]. So far only a handful of STAT5B inhibitor chemotypes were described [114], which primarily act by interfering with STAT5B dimer formation via binding to its SH2 domain [115]. In the above mentioned example, the authors have used multiple scoring functions (molecular mechanics energy, surface area and others) of the FlexX docking software to select 30 virtual hits by consensus, whose testing ultimately resulted in the micromolar inhibitor IST5-002.

Focused screening

In focused screening, computational approaches are applied to cut back on the demand of experimental testing by compiling focused compound libraries from the physically (or commercially) available compound pool. This includes various possibilities by mostly ligand-based (and occasionally protein-based) screening methods, or their combinations.

A typical area for the use of focused libraries is kinase drug discovery. Protein kinases are a family of over 600 human enzymes [116], whose core function is intracellular signaling, through the phosphorylation of specific sidechains of protein substrates [117]. Their key role in signaling eventually nominates most kinases as a therapeutic target against some form of disease, mostly oncological and autoimmune disorders that stem from abnormal signaling [118]. Since the source of phosphoryl groups is uniformly the adenosine triphosphate (ATP) molecule, the most important concept for kinase inhibition is the use of ATP-competitive small molecules – although other approaches also exist [119]. ATP-competitive binders are well-characterized from a structural/topological point of view, therefore several approaches are available to compile kinase-focused molecular libraries. These include descriptor-based approaches relying on multi-criteria optimization rules [16] or the statistical characterization of kinase-like descriptor space [120], as well as fragment-based approaches based on the characteristic molecular recognition patterns between small-molecule kinase inhibitors and the conserved hinge region of kinases [121]. The concept was successfully applied in virtual screening campaigns against diverse kinase targets, such as Janus kinases (JAK) [122, 123] or inositol phosphate kinases [124], and such libraries are offered commercially by several compound vendors. A collection of guidelines and protocols for kinase library design was published recently [125].

Focused libraries can be compiled based on more target-specific concepts as well. In another effort to discover inhibitors of the above mentioned STAT5 transcription factor, Natarajan and colleagues have compiled a library of virtually O-phosphorylated natural products [126]. Here, the authors have exploited the easy synthetic access to O-phosphorylated natural products, as well as the heavily conserved, core recognition motif between the Arg618 sidechain of the SH2 domain binding site and the phosphotyrosine moiety of the STAT5 substrate sequence. They have collected a library of suitable natural products and implemented an algorithmic step to modify their SMILES strings by adding an O-phosphoryl group to their phenolic moieties. The resulting virtual compounds were used for docking, and nine selected virtual hits were synthesized by a two-step phosphorylation/debenzylation process starting from the unphosphorylated, commercially available analogs. The work has ultimately resulted in Stafia-1, the first inhibitor of STAT5A with selectivity over its close homologue STAT5B [126].

In the traditional sense, focused libraries aim to provide enhanced hit rates against a specific target or target class, as seen above. However, the same methodologies can be customized and applied for an inverse concept, in which the goal is to provide a sort of “general utility” compound library, which would maximize the chances of finding hits against any protein target, by the experimental screening of a strongly limited number of compounds. Evidently, this idea is most feasible for the design of fragment libraries, due to the more manageable size of the fragment chemical space. Pharmacophores are ideal representations of the structural diversity of the 3D pattern of interacting features (H-bond donors, acceptors, ionic centers, etc.) that are present in a fragment-sized molecule. We bring two examples for recent works that have applied the concept of pharmacophores for fragment library design. In the F2X libraries introduced by Wollenhaupt et al. for crystallographic fragment screening, commercially

available fragments were clustered based on their 3D shape and pharmacophore similarity calculated with the ROCS (Rapid Overlay of Chemical Structures) method, to yield a small library of representative fragments with diverse pharmacophore arrangements [127]. The method was validated against two targets with excellent hit rates. In the meantime, we have developed a library design method (SpotXplorer) with a different concept: we have collected experimentally validated pharmacophore arrangements from publicly available fragment-protein complexes and compiled a fragment library with maximized coverage of the representative 2- and 3-point fragment pharmacophores [10]. The resulting library of 96 fragments have yielded diverse hits against several classes of target proteins, including the two SARS-CoV-2 protein targets. In both cases, the objective was to reach diversity within the library, as opposed to the above examples where an overall similarity of the compounds is implied, on the level of descriptors or substructures.

Sequential screening

Sequential screening aims to balance the cost vs. accuracy of computational methods, or cost vs. relevance of experimental screening methods. Sequential workflows apply gradually more expensive and more accurate steps to a gradually smaller pool of compounds, eliminating a large portion of the screening deck at each step, and progressing a fraction of the best compounds into the next step.

Our recent work on SETD2 inhibitors nicely illustrates the concept [51]. SETD2 is a histone methyltransferase that was recently established as a new oncotarget, primarily against certain types of leukemia [128, 129]. So far, only a handful of nucleoside analog natural products were reported as SETD2 inhibitors, which prompted us to explore a wider chemical space in search of new compounds. Our hit discovery workflow started with a commercial library of close to 6 million compounds and included gradually more accurate screening methods to cut down on computing time, while retaining the best candidates for the next step (Figure 3). This started with a PAINS filtering step to remove interference compounds (cf. section 2), then a general and permissive substructural pattern (extracted from known inhibitors) was used to shrink the compound pool by an order of magnitude, followed by a pharmacophore screening step to the same effect. Ligand docking itself has further reduced the number of compounds, from which the best unique molecules were visually inspected for their predicted binding poses, and clustered based on their chemical diversity. We should note that the calculations were performed on a single workstation with 8 CPUs in approx. 5 days; without the early filtering steps, the docking step would have taken years on the same infrastructure. The stepwise logic was kept in experimental testing as well: here, the selected 22 virtual hits were first evaluated in an enzyme-based assay format and only the three best compounds were progressed to the more expensive and laborious cell-based characterization. One of the compounds clearly stood out and was reported as the primary hit of this work.

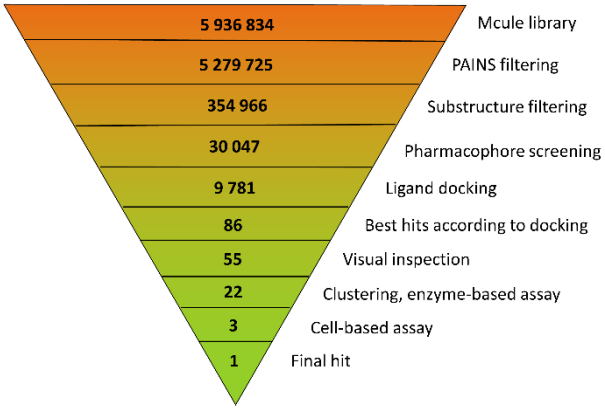

Figure 3. Stepwise screening workflow to discover new SETD2 inhibitors. From each step, a smaller fraction of the compound pool is progressed to the next step. (The figure was adapted as a modified version of Figure 3A from our recent work [51], © 2021 by the authors, licensee MDPI, Basel, Switzerland, under the Creative Commons Attribution (CC BY 4.0) license.)

Other works employ stepwise protocols with different screening steps, for example Castillo-González et al. have included rigorously developed QSAR models as an early filtering step in the discovery of G-quadruplex (G4) stabilizer ligands [130], reporting an excellent hitrate of 23.5%. Similarly, a COVID-19 drug repurposing study has demonstrated the utility of shape similarity as an early filtering step, and ultimately resulted in the discovery of several existing drugs with anti-SARS-CoV-2 activity in human lung cells [131]. The steps of a sequential workflow can represent more subtle methodological differences. The recent work of Zhang et al. employed a sequential workflow of different docking steps to discover new inhibitors of STAT3, another member of the STAT transcription factor family mentioned above [132]. A commercial compound library was docked to the STAT3 binding site first by DOCK 4.0 [133], then the top 10% was promoted to a second round of docking by Glide SP (single precision mode) [134, 135], then the top 500 molecules were docked again by Glide XP (extra precision mode) [136], representing gradually more sophisticated docking steps. After the experimental testing of 100 molecules, the work has resulted in the discovery of a series of benzothiazole-based STAT3 inhibitors.

In experimental screening, a sequential workflow is generally used to minimize costs. While screening larger libraries, it is customary to carry out a primary screening with one concentration of the candidate compounds (in duplicate in triplicate) and then advance the primary hits to more thorough characterization by measuring their dose-response curves or promoting them into higher-level, mostly cell-based biochemical experiments. This is considered a default practice and will not be further illustrated here.

Iterative and integrated screening

Iterative screening follows the same philosophy as sequential screening, with the key difference that in iterative screening, results from a more advanced step are fed back to an earlier step, in order to increase overall performance and pick up candidates that might have been left out in previous iterations. In terms of virtual screening, a typical example of this is the Deep Docking workflow mentioned above, where “true” docking scores are fed back to the neural network in order to enhance the predictive capability and thereby the overall performance of the deep learning model [61]. The improved model is then used to nominate another iteration of compounds into the docking step.

In the interplay of virtual and experimental screening, the experimental results that are generated from the compounds selected in the first round of virtual screening can be used to iteratively refine the VS methodology and promote new hit compounds in subsequent iterations. The VS methodology may be the same, or different, than in the earlier iterations. While some authors use the terms “iterative” and “integrated” interchangeably, others distinguish between the two by specifying that in the latter, the computational and experimental methodologies are integrated more closely within the hit retrieval process. One example of an integrated workflow would be to train the primary virtual screening methodology (e.g. a 2D similarity-based method or a machine learning model) with existing experimental data, and plugging in the newly generated experimental results into later iterations. In fact, a recent methodological paper by Miyao and Funatsu provides a detailed, retrospective comparison of available machine learning methods for iterative screening [137]. Since the advantages of the past decade have brought forth endless possibilities for the interplay of various computational and experimental methods, we feel that the line between these terms was blurred, and therefore we will not discuss these approaches separately.

In a recent contribution, the authors compare several approaches to implement 2D and 3D virtual screening methods into integrated workflows [138]. In their recurring approach, a 2D similarity-based method is trained on existing active compounds and employed as a first round of virtual screening. Hits are selected and the results of the experimental testing are used to train a 3D ligand-based (e.g. pharmacophore screening) or structure-based (e.g. docking) virtual screening method to retrieve a second round of hits for testing. This approach was demonstrated on several targets, including the serine protease C1s in a recent example [139].

In a paper by Merck, the authors provide a comprehensive overview of integrated virtual screening strategies, illustrated on three case studies, corresponding to different scenarios regarding the availability of target structures and known active ligands [140]. Their third case study details the discovery of tool compounds of a new, non-disclosed immuno-oncological target with no known modulators, or structural information at the time of the study. An initial set of actives was retrieved by two parallel approaches: i) QSAR models based on the known ligands of homologous (25-30% sequence identity) target family members, and ii) a pharmacophore model based on known enzymatic substrates of Target A. The initial biochemical screen has confirmed 25 and 33 active hits from a set of 4,000 virtual hits (2,000 for each arm). The activity information of these hits, as well as the inactives, were utilized to conduct four rounds of iterative focused screening (IFS) – a technique described by Merck for the CNS target alanine–serine–cysteine transporter-1 (Asc-1) [141]. Briefly, the information from the previous iteration is fed into three complementary methods (2D QSAR, 3D shape screening, activity fingerprints) to retrieve both similar, and structurally distinct additional hits. Finally, the study resulted in 19 distinct hit series for a previously underexplored target in an impressive timeframe of 6 months.

5. Conclusion

Both virtual and experimental screening are indispensable methodologies in the toolbox of early phase drug discovery. We have reviewed the current state-of-the-art in virtual and experimental screening, concentrating on the advances from the last decade. A main section was dedicated to illustrating the main options for the integration of these methodologies in order to increase the success rate and cost efficiency of hit discovery. Notably, we find that the field has undergone major paradigm shifts in the last ten years. The quick expansion of the accessible chemical space necessitates new algorithmic concepts and smarter integration of screening

methods to discover and utilize novel chemical structures. We envision an even tighter integration of the available methodological tools over time, as explained in more detail below.

Expert Opinion

Recent studies revealed that experimental and virtual screening contributed to the identification of clinical candidates for more than half of the successful preclinical programs. Although integrative approaches, such as focused screens were used successfully, more efficient iterative protocols utilize the knowledge generated at the repeated cycles of virtual and experimental screens. In this scenario, large chemical space can be first focused by virtual screening approaches and the resulted subset becomes feasible for experimental testing. The results are then available to improve the predictive power of virtual screening tools. Artificial intelligence methods can be effectively used to extract and implement this target-specific knowledge that can be used to extend the chemical space of previous hits. Experimental screening of this library would provide higher hit rates and more diverse chemotypes with better profile.

During the recent years, a significantly larger part of the chemical space became available. In addition to physical compound collections, large virtual compound databases are generated using commercially available building blocks and defined sets of robust reactions. Multistep virtual synthesis provided large synthetically accessible libraries due to the combinatorial explosion. Efficient extraction of synthetic knowledge from laboratory E-notebooks and high throughput experimentation used for reaction optimization would provide more and more robust reactions. Together with the ever-increasing set of available building blocks, these trends support the increased coverage of the chemical space. Efficient virtual screening of these libraries needs extended computational capacity and improved algorithms. Virtual screening benefits well from the application of GPU clusters and cloud architectures. Better access to these hardware resources and further improvements in software parallelization, however, would be needed to extend the present boundaries. More effective searching in extremely large databases would require new virtual screening methods in both algorithms and implementation.

Reviewing the recent improvements in experimental screening, we mention only two major goals. Quantitative HTS techniques provide more reliable experimental datasets and increase the confidence in hit prioritization. Better quality data, however, improves the iterative learning curve of computational tools that can be used for virtual screening. This advantage would be effectively realized in integrated experimental and virtual screening campaigns, particularly in iterative setups. The other important achievement in experimental screening is the increasing availability of DNA encoded libraries (DELs). Although the technology is known for more than a decade, dramatic improvements in DNA-compatible chemistries resulted in a wide range of large DEL collections. Simultaneous developments in DEL selection techniques made DEL screening available even in living cells. An important aspect of DEL technologies is the democratization of large-scale screening that is now available even in academia. DEL screening provides large, annotated datasets of actives and inactives that might be considered as input to classification techniques. Developing such tools would be only the first step to utilize DEL datasets for virtual screening. Artificial intelligence tools, however, would extend their scope further and would connect the DEL space to the synthetically accessible chemical space. Considering the differences between the current set of robust reactions and the actual set of DNA-compatible reactions, both fields would benefit from further synthetic research.

Acknowledgement

This work was supported under project No. OTKA K 135150 by the Ministry of Innovation and Technology of Hungary, from the National Research, Development and Innovation Fund, financed under the K type funding scheme. The work of D.B. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the ÚNKP-21-5 New National Excellence Program of the Ministry for Innovation and Technology.

References

[1] Keserű, G. M.; Makara, G. M. The Influence of Lead Discovery Strategies on the Properties of Drug Candidates. *Nat. Rev. Drug Discov.*, **2009**, *8* (3), 203–212. https://doi.org/10.1038/nrd2796.

[2] Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *J. Med. Chem.*, **2018**, *61* (21), 9442–9468. https://doi.org/10.1021/ACS.JMEDCHEM.8B00675.

[3] Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discov.*, **2002**, *1* (11), 882–894. https://doi.org/10.1038/nrd941.

[4] Polgár, T.; Keserű, G. M. Integration of Virtual and High Throughput Screening in Lead Discovery Settings. *Comb. Chem. High Throughput Screen.*, **2011**, *14* (10), 889–897. https://doi.org/10.2174/138620711797537148.

[5] Tanrikulu, Y.; Krüger, B.; Proschak, E. The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discov. Today*, **2013**, *18* (7–8), 358–364. https://doi.org/10.1016/J.DRUDIS.2013.01.007.

**** A review paper on parallel, iterative and integrated VS workflows up until 2013.**

[6] Willems, H.; De Cesco, S.; Svensson, F. Computational Chemistry on a Budget: Supporting Drug Discovery with Limited Resources. *J. Med. Chem.*, **2020**, *63* (18), 10158–10169. https://doi.org/10.1021/ACS.JMEDCHEM.9B02126.

[7] Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual Screening Web Servers: Designing Chemical Probes and Drug Candidates in the Cyberspace. *Brief. Bioinform.*, **2021**, *22* (2), 1790–1818. https://doi.org/10.1093/BIB/BBAA034.

[8] Bajusz, D.; Rácz, A.; Héberger, K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In *Comprehensive Medicinal Chemistry III*; Chackalamannil, S., Rotella, D. P., Ward, S. E., Eds.; Elsevier: Oxford, 2017; pp 329–378. https://doi.org/10.1016/B978-0-12-409547-2.12345-5.

[9] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.*, **1997**, *23* (1–3), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

[10] Bajusz, D.; Wade, W. S.; Satała, G.; Bojarski, A. J.; Ilaš, J.; Ebner, J.; Grebien, F.; Papp, H.; Jakab, F.; Douangamath, A.; et al. Exploring Protein Hotspots by Optimized Fragment Pharmacophores. *Nat. Commun.*, **2021**, *12*, 3201. https://doi.org/10.1038/s41467-021-23443-y.

[11] Murray, C. W.; Rees, D. C. Opportunity Knocks: Organic Chemistry for Fragment-Based Drug Discovery (FBDD). *Angew. Chemie Int. Ed.*, **2016**, *55* (2), 488–492. https://doi.org/10.1002/anie.201506783.

[12] Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.*, **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.

*** A landmark paper that introduces the concept of PAINS, currently the most widespread ruleset to exclude interference compounds from screening libraries.**

[13] Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K. M.; Schepartz, A.; Wang, S. The Ecstasy and Agony of Assay Interference Compounds. *J. Med. Chem.*, **2017**, *60* (6), 2165–2168. https://doi.org/10.1021/ACS.JMEDCHEM.7B00229.

[14] Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature*, **2014**, *513* (7519), 481–483. https://doi.org/10.1038/513481a.

[15] Kelemen, Á. A.; Ferenczy, G. G.; Keserű, G. M. A Desirability Function-Based Scoring Scheme for Selecting Fragment-like Class A Aminergic GPCR Ligands. *J. Comput. Aided. Mol. Des.*, **2015**, *29* (1), 59–66. https://doi.org/10.1007/s10822-014-9804-5.

[16] Bajusz, D.; Ferenczy, G. G.; Keserű, G. M. Property-Based Characterization of Kinase-like Ligand Space for Library Design and Virtual Screening. *Medchemcomm*, **2015**, *6* (10), 1898–1904. https://doi.org/10.1039/C5MD00253B.

[17] Nicolaou, C. A.; Brown, N. Multi-Objective Optimization Methods in Drug Design. *Drug Discov. Today Technol.*, **2013**, *10* (3), e427–e435. https://doi.org/10.1016/J.DDTEC.2013.02.001.

[18] Sánchez-Rodríguez, A.; Pérez-Castillo, Y.; Schürer, S. C.; Nicolotti, O.; Mangiatordi, G. F.; Borges, F.; Cordeiro, M. N. D. S.; Tejera, E.; Medina-Franco, J. L.; Cruz-Monteagudo, M. From Flamingo Dance to (Desirable) Drug Discovery: A Nature-Inspired Approach. *Drug Discov. Today*, **2017**, *22* (10), 1489–1502. https://doi.org/10.1016/J.DRUDIS.2017.05.008.

[19] Lambrinidis, G.; Tsantili-Kakoulidou, A. Challenges with Multi-Objective QSAR in Drug Discovery. *Expert Opinion on Drug Discovery*. Taylor & Francis September 2, 2018, pp 851–859. https://doi.org/10.1080/17460441.2018.1496079.

[20] Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. 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.*, **1999**, *1* (1), 55–68. https://doi.org/10.1021/CC9800071.

[21] Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.*, **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

[22] Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chemie Int. Ed.*, **1999**, *38* (24), 3743–3748.

[23] Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-Based Lead Discovery: Leads by Design. *Drug Discov. Today*, **2005**, *10* (14), 987–992.

[24] Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "Rule of Three" for Fragment-Based Lead Discovery? *Drug Discov. Today*, **2003**, *8* (19), 876–877. https://doi.org/10.1016/S1359-6446(03)02831-9.

[25] Walters, W. P.; Namchuk, M. Designing Screens: How to Make Your Hits a Hit. *Nat.*

Rev. Drug Discov., **2003**, *2* (4), 259–266. https://doi.org/10.1038/nrd1063.

[26] Bohacek, R. S.; Mcmartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.*, **1996**, *16* (1), 3–50.

[27] Dobson, C. M. Chemical Space and Biology. *Nature*, **2004**, *432*, 824–828. https://doi.org/10.1038/NATURE03192.

[28] Van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.*, **2019**, *59* (2), 644–651. https://doi.org/10.1021/ACS.JCIM.8B00737.

*** A perspective on the use of large virtual libraries, with a collection of publicly available virtual libraries.**

[29] REAL Database - Enamine https://enamine.net/compound-collections/real-compounds/real-database (accessed Jan 14, 2022).

[30] Mcule - Ultimate Database Project https://ultimate.mcule.com/ (accessed Jan 14, 2022).

[31] Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.*, **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/ACS.JCIM.0C00675.

[32] Boehm, M.; Wu, T. Y.; Haussen, H.; Lemmen, C. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces. *J. Med. Chem.*, **2008**, *51* (8), 2468–2480. https://doi.org/10.1021/JM0707727.

[33] Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today*, **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/J.DRUDIS.2019.02.013.

[34] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.*, **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

[35] PRACE - Partnership for Advanced Computing in Europe https://prace-ri.eu/ (accessed Jan 17, 2022).

[36] Santos-Martins, D.; Solis-Vasquez, L.; Tillack, A. F.; Sanner, M. F.; Koch, A.; Forli, S. Accelerating AutoDock4 with GPUs and Gradient-Based Local Search. *J. Chem. Theory Comput.*, **2021**, *17* (2), 1060–1073. https://doi.org/10.1021/ACS.JCTC.0C01006.

[37] Acharya, A.; Agarwal, R.; Baker, M. B.; Baudry, J.; Bhowmik, D.; Boehm, S.; Byler, K. G.; Chen, S. Y.; Coates, L.; Cooper, C. J.; et al. Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J. Chem. Inf. Model.*, **2020**, *60* (12), 5832–5852. https://doi.org/10.1021/ACS.JCIM.0C01010.

[38] Bender, B. J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C. M.; Stein, R. M.; Fink, E. A.; Balius, T. E.; Carlsson, J.; Irwin, J. J.; et al. A Practical Guide to Large-Scale Docking. *Nat. Protoc.*, **2021**, *16* (10), 4799–4832. https://doi.org/10.1038/s41596-021-00597-z.

**** An instructive paper on conducting large-scale virtual screens with the DOCK software.**

[39] FastROCS Toolkit | Real-Time Shape Similarity | Lead Discovery https://www.eyesopen.com/molecular-modeling-fastrocs (accessed Jan 17, 2022).

[40] Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.*, **1996**, *17* (14), 1653–1666.

[41] Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.*, **2005**, *48* (5), 1489–1495. https://doi.org/10.1021/jm040163o.

[42] Koes, D. R.; Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *J. Chem. Inf. Model.*, **2011**, *51* (6), 1307–1314. https://doi.org/10.1021/ci200097m.

[43] Koes, D. R.; Camacho, C. J. ZINCPharmer: Pharmacophore Search of the ZINC Database. *Nucleic Acids Res.*, **2012**, *40* (Web Server issue), W409-14. https://doi.org/10.1093/nar/gks378.

[44] Lowe, D. Virtual Screening Versus the Numbers https://www.science.org/content/blog-post/virtual-screening-versus-numbers (accessed Feb 22, 2022).

[45] Gorgulla, C.; Boeszoermenyi, A.; Wang, Z. F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature*, **2020**, *580*, 663–668. https://doi.org/10.1038/s41586-020-2117-z.

**** A paper introducing VirtualFlow, a flexible software platform for ultra-large virtual screens.**

[46] Murugan, N. A.; Podobas, A.; Gadioli, D.; Vitali, E.; Palermo, G.; Markidis, S. A Review on Parallel Virtual Screening Softwares for High-Performance Computers. *Pharmaceuticals*, **2022**, *15* (1), 63. https://doi.org/10.3390/PH15010063.

[47] Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature*, **2019**, *566*, 224–229. https://doi.org/10.1038/s41586-019-0917-9.

[48] Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; et al. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature*, **2021**, 1–8. https://doi.org/10.1038/s41586-021-04220-9.

[49] Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/ACS.JCIM.5B00559.

[50] Luttens, A.; Gullberg, H.; Abdurakhmanov, E.; Vo, D. D.; Akaberi, D.; Talibov, V. O.; Nekhotiaeva, N.; Vangeel, L.; Jonghe, S. De; Jochmans, D.; et al. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J. Am. Chem. Soc.*, **2022**, jacs.1c08402. https://doi.org/10.1021/JACS.1C08402.

[51] Bajusz, D.; Bognár, Z.; Ebner, J.; Grebien, F.; Keserű, G. M. Discovery of a Non-Nucleoside SETD2 Methyltransferase Inhibitor against Acute Myeloid Leukemia. *Int. J. Mol. Sci.*, **2021**, *22* (18), 10055. https://doi.org/10.3390/ijms221810055.

[52] Rácz, A.; Palkó, R.; Csányi, D.; Riedl, Z.; Bajusz, D.; Keserű, G. M. Consensus Virtual Screening Identified [1,2,4]Triazolo[1,5-b]Isoquinolines As MELK Inhibitor Chemotypes. *ChemMedChem*, **2022**, *17* (2), e202100569. https://doi.org/10.1002/cmdc.202100569.

[53] Perez-Castillo, Y.; Sotomayor-Burneo, S.; Jimenes-Vargas, K.; Gonzalez-Rodriguez,

M.; Cruz-Monteagudo, M.; Armijos-Jaramillo, V.; Cordeiro, M. N. D. S.; Borges, F.; Sánchez-Rodríguez, A.; Tejera, E. CompScore: Boosting Structure-Based Virtual Screening Performance by Incorporating Docking Scoring Function Components into Consensus Scoring. *J. Chem. Inf. Model.*, **2019**, *59* (9), 3655–3666. https://doi.org/10.1021/ACS.JCIM.9B00343/SUPPL_FILE/CI9B00343_SI_001.PDF.

[54] Perez-Castillo, Y.; Morales Helguera, A.; Natália, M.; Cordeiro, D. S.; Tejera, E.; Paz-Y-Miño, C.; Sánchez-Rodríguez, A.; Borges, F.; Cruz-Monteagudo, M. Fusing Docking Scoring Functions Improves the Virtual Screening Performance for Discovering Parkinson's Disease Dual Target Ligands. *Curr. Neuropharmacol.*, **2017**, *15* (8), 1107–1116. https://doi.org/10.2174/1570159X15666170109143757.

[55] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, **2021**, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[56] Method of the Year 2021: Protein Structure Prediction. *Nat. Methods*, **2022**, *19*, 1. https://doi.org/10.1038/s41592-021-01380-4.

[57] Rácz, A.; Bajusz, D.; Miranda-Quintana, R. A.; Héberger, K. Machine Learning Models for Classification Tasks Related to Drug Safety. *Mol. Divers.*, **2021**, *25*, 1409–1424. https://doi.org/10.1007/s11030-021-10239-x.

[58] Kimber, T. B.; Chen, Y.; Volkamer, A. Deep Learning in Virtual Screening: Recent Applications and Developments. *Int. J. Mol. Sci.*, **2021**, *22*, 4435. https://doi.org/10.3390/IJMS22094435.

[59] Ahmed, L.; Georgiev, V.; Capuccini, M.; Toor, S.; Schaal, W.; Laure, E.; Spjuth, O. Efficient Iterative Virtual Screening with Apache Spark and Conformal Prediction. *J. Cheminform.*, **2018**, *10* (1), 1–8. https://doi.org/10.1186/S13321-018-0265-Z.

[60] Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening through Molecular Pool-Based Active Learning. *Chem. Sci.*, **2021**, *12* (22), 7866–7881. https://doi.org/10.1039/D0SC06805E.

[61] Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A. T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.*, **2020**, *6* (6), 939–949. https://doi.org/10.1021/ACSCENTSCI.0C00229.

*** First publication on the concept of Deep Docking.**

[62] Gentile, F.; Fernandez, M.; Ban, F.; Ton, A. T.; Mslati, H.; Perez, C. F.; Leblanc, E.; Yaacoub, J. C.; Gleave, J.; Stern, A.; et al. Automated Discovery of Noncovalent Inhibitors of SARS-CoV-2 Main Protease by Consensus Deep Docking of 40 Billion Small Molecules. *Chem. Sci.*, **2021**, *12* (48), 15960–15974. https://doi.org/10.1039/D1SC05579H.

[63] Gentile, F.; Yaacoub, J. C.; Gleave, J.; Fernandez, M.; Ton, A.-T.; Ban, F.; Stern, A.; Cherkasov, A. Artificial Intelligence–Enabled Virtual Screening of Ultra-Large Chemical Libraries with Deep Docking. *Nat. Protoc.*, **2022**. https://doi.org/10.1038/s41596-021-00659-2.

**** An instructive paper for conducting Deep Docking calculations on large libraries.**

[64] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.*, **2004**, *47* (12), 2977–2980. https://doi.org/10.1021/jm030580l.

[65] Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A Soft Tool for 3D Drug Design of Protein Targets by Artificial Intelligence and Classical Algorithm. *Brief. Bioinform.*, **2021**, *22* (3). https://doi.org/10.1093/bib/bbaa161.

[66] Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.*, **2013**, *13* (14), 1713–1741. https://doi.org/10.2174/1568026611313140011.

[67] Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva Caracuel, E.; González-Díaz, H. PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives. *ACS Comb. Sci.*, **2020**, *22* (3), 129–141. https://doi.org/10.1021/acscombsci.9b00166.

[68] Kleandrova, V. V.; Speck-Planche, A. PTML Modeling for Alzheimer's Disease: Design and Prediction of Virtual Multi-Target Inhibitors of GSK3B, HDAC1, and HDAC6. *Curr. Top. Med. Chem.*, **2020**, *20* (19), 1661–1676. https://doi.org/10.2174/1568026620666200607190951.

[69] Kleandrova, V. V.; Scotti, M. T.; Scotti, L.; Speck-Planche, A. Multi-Target Drug Discovery via PTML Modeling: Applications to the Design of Virtual Dual Inhibitors of CDK4 and HER2. *Curr. Top. Med. Chem.*, **2021**, *21* (7), 661–675. https://doi.org/10.2174/1568026621666210119112845.

[70] Kleandrova, V. V.; Speck-Planche, A. PTML Modeling for Pancreatic Cancer Research: In Silico Design of Simultaneous Multi-Protein and Multi-Cell Inhibitors. *Biomedicines*, **2022**, *10* (2), 491. https://doi.org/10.3390/biomedicines10020491.

[71] Kleandrova, V. V.; Scotti, M. T.; Speck-Planche, A. Computational Drug Repurposing for Antituberculosis Therapy: Discovery of Multi-Strain Inhibitors. *Antibiotics*, **2021**, *10* (8), 1005. https://doi.org/10.3390/antibiotics10081005.

[72] Diéguez-Santana, K.; Casañola-Martin, G. M.; Green, J. R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Curr. Top. Med. Chem.*, **2021**, *21* (9), 819–827. https://doi.org/10.2174/1568026621666210331161144.

[73] Wu, P.; Noland, C.; Ultsch, M.; Edwards, B.; Harris, D.; Mayer, R.; Harris, S. F. Developments in the Implementation of Acoustic Droplet Ejection for Protein Crystallography. *J. Lab. Autom.*, **2016**, *21* (1), 97–106. https://doi.org/10.1177/2211068215598938.

[74] Chodera, J.; Lee, A. A.; London, N.; von Delft, F. Crowdsourcing Drug Discovery for Pandemics. *Nat. Chem.*, **2020**, *12* (7), 581. https://doi.org/10.1038/s41557-020-0496-2.

*** A brief correspondence, introducing the concept of crowdsourced drug discovery through the COVID Moonshot collaboration.**

[75] Douangamath, A.; Fearon, D.; Gehrtz, P.; Krojer, T.; Lukacik, P.; Owen, C. D.; Resnick, E.; Strain-Damerell, C.; Aimon, A.; Ábrányi-Balogh, P.; et al. Crystallographic and Electrophilic Fragment Screening of the SARS-CoV-2 Main Protease. *Nat. Commun.*, **2020**, *11*, 5047. https://doi.org/10.1038/s41467-020-18709-w.

[76] Schuller, M.; Correy, G. J.; Gahbauer, S.; Fearon, D.; Wu, T.; Díaz, R. E.; Young, I. D.; Carvalho Martins, L.; Smith, D. H.; Schulze-Gahmen, U.; et al. Fragment Binding to the Nsp3 Macrodomain of SARS-CoV-2 Identified through Crystallographic Screening and Computational Docking. *Sci. Adv.*, **2021**, *7* (16), eabf8711. https://doi.org/10.1126/sciadv.abf8711.

[77] Resnick, E.; Bradley, A.; Gan, J.; Douangamath, A.; Krojer, T.; Sethi, R.; Geurink, P. P.; Aimon, A.; Amitai, G.; Bellini, D.; et al. Rapid Covalent-Probe Discovery by Electrophile-Fragment Screening. *J. Am. Chem. Soc.*, **2019**, *141* (22), 8951–8968. https://doi.org/10.1021/JACS.9B02822/SUPPL_FILE/JA9B02822_SI_004.PDF.

[78] Wigglesworth, M.; Murray, D. Leveraging the Power of HTS in the New and Rapidly Evolving Collaborative Drug Discovery Landscape. *Drug Target Review*. 2018.

[79] Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of Commercially Available Compounds for HTS. *Drug Discov. Today*, **2019**, *24* (2), 390–402. https://doi.org/10.1016/J.DRUDIS.2018.10.016.

[80] Zeng, W.; Guo, L.; Xu, S.; Chen, J.; Zhou, J. High-Throughput Screening Technology in Industrial Biotechnology. *Trends Biotechnol.*, **2020**, *38* (8), 888–906. https://doi.org/10.1016/J.TIBTECH.2020.01.001.

[81] Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening: A Titration-Based Approach That Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci.*, **2006**, *103* (31), 11473–11478. https://doi.org/10.1073/PNAS.0604348103.

[82] Tice, R. R.; Austin, C. P.; Kavlock, R. J.; Bucher, J. R. Improving the Human Hazard Characterization of Chemicals: A Tox21 Update. *Environ. Health Perspect.*, **2013**, *121* (7), 756–765. https://doi.org/10.1289/EHP.1205784.

[83] Shockley, K. R. Quantitative High-Throughput Screening Data Analysis: Challenges and Recent Advances. *Drug Discov. Today*, **2015**, *20* (3), 296–300. https://doi.org/10.1016/J.DRUDIS.2014.10.005.

[84] Shockley, K. R.; Gupta, S.; Harris, S. F.; Lahiri, S. N.; Peddada, S. D. Quality Control of Quantitative High Throughput Screening Data. *Front. Genet.*, **2019**, *10* (MAY), 387. https://doi.org/10.3389/FGENE.2019.00387.

[85] Swinney, D. C.; Lee, J. A. Recent Advances in Phenotypic Drug Discovery. *F1000Research*, **2020**, *9*, 944. https://doi.org/10.12688/F1000RESEARCH.25813.1.

**** A recent review on the challenges and progress in phenotypic drug discovery.**

[86] Eder, J.; Sedrani, R.; Wiesmann, C. The Discovery of First-in-Class Drugs: Origins and Evolution. *Nat. Rev. Drug Discov.*, **2014**, *13* (8), 577–587. https://doi.org/10.1038/nrd4336.

[87] Naryshkin, N. A.; Weetall, M.; Dakka, A.; Narasimhan, J.; Zhao, X.; Feng, Z.; Ling, K. K. Y.; Karp, G. M.; Qi, H.; Woll, M. G.; et al. SMN2 Splicing Modifiers Improve Motor Function and Longevity in Mice with Spinal Muscular Atrophy. *Science (80-.).*, **2014**, *345* (6197), 688–693. https://doi.org/10.1126/SCIENCE.1250127.

[88] Van Goor, F.; Hadida, S.; Grootenhuis, P. D. J.; Burton, B.; Cao, D.; Neuberger, T.; Turnbull, A.; Singh, A.; Joubran, J.; Hazlewood, A.; et al. Rescue of CF Airway Epithelial Cell Function in Vitro by a CFTR Potentiator, VX-770. *Proc. Natl. Acad. Sci.*, **2009**, *106* (44), 18825–18830. https://doi.org/10.1073/PNAS.0904709106.

[89] Lemm, J. A.; O'Boyle, D.; Liu, M.; Nower, P. T.; Colonno, R.; Deshpande, M. S.; Snyder, L. B.; Martin, S. W.; St. Laurent, D. R.; Serrano-Wu, M. H.; et al. Identification of Hepatitis C Virus NS5A Inhibitors. *J. Virol.*, **2010**, *84* (1), 482–491. https://doi.org/10.1128/JVI.01360-09.

[90] Kenny, H. A.; Lal-Nag, M.; White, E. A.; Shen, M.; Chiang, C. Y.; Mitra, A. K.; Zhang, Y.; Curtis, M.; Schryver, E. M.; Bettis, S.; et al. Quantitative High Throughput Screening Using a Primary Human Three-Dimensional Organotypic Culture Predicts in Vivo Efficacy. *Nat. Commun.*, **2015**, *6*, 6220. https://doi.org/10.1038/ncomms7220.

[91] Aulner, N.; Danckaert, A.; Ihm, J. E.; Shum, D.; Shorte, S. L. Next-Generation Phenotypic Screening in Early Drug Discovery for Infectious Diseases. *Trends Parasitol.*, **2019**, *35* (7), 559–570. https://doi.org/10.1016/J.PT.2019.05.004.

[92] Riva, L.; Yuan, S.; Yin, X.; Martin-Sancho, L.; Matsunaga, N.; Pache, L.; Burgstaller-Muehlbacher, S.; De Jesus, P. D.; Teriete, P.; Hull, M. V.; et al. Discovery of SARS-CoV-2 Antiviral Drugs through Large-Scale Compound Repurposing. *Nature*, **2020**, *586*, 113–119. https://doi.org/10.1038/s41586-020-2577-1.

*** An important application of phenotypic drug discovery, showcasing its utility in drug repurposing against an ongoing pandemic.**

[93] Haasen, D.; Schopfer, U.; Antczak, C.; Guy, C.; Fuchs, F.; Selzer, P. How Phenotypic Screening Influenced Drug Discovery: Lessons from Five Years of Practice. *Assay Drug Dev. Technol.*, **2017**, *15* (6), 239–246. https://doi.org/10.1089/ADT.2017.796.

[94] Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.*, **2017**, *16* (8), 531–543. https://doi.org/10.1038/nrd.2017.111.

[95] Corrò, C.; Novellasdemunt, L.; Li, V. S. W. Making Cell Culture More Physiological: A Brief History of Organoids. *Am. J. Physiol. - Cell Physiol.*, **2020**, *319* (1), C151. https://doi.org/10.1152/AJPCELL.00120.2020.

[96] Kim, J.; Koo, B. K.; Knoblich, J. A. Human Organoids: Model Systems for Human Biology and Medicine. *Nat. Rev. Mol. Cell Biol.*, **2020**, *21* (10), 571–584. https://doi.org/10.1038/s41580-020-0259-3.

[97] Mattiazzi Usaj, M.; Styles, E. B.; Verster, A. J.; Friesen, H.; Boone, C.; Andrews, B. J. High-Content Screening for Quantitative Cell Biology. *Trends Cell Biol.*, **2016**, *26* (8), 598–611. https://doi.org/10.1016/J.TCB.2016.03.008.

[98] Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol. Transl. Sci.*, **2021**, *4*, 1265–1279. https://doi.org/10.1021/ACSPTSCI.1C00118.

**** A Comprehensive Review with Succesful Stories and Future Challenges: A seminal review on the general methodology and application of DNA-encoded libraries.**

[99] Brenner, S.; Lerner, R. A. Encoded Combinatorial Chemistry. *Proc. Natl. Acad. Sci.*, **1992**, *89* (12), 5381–5383. https://doi.org/10.1073/PNAS.89.12.5381.

[100] Halford, B. How DNA-Encoded Libraries Are Revolutionizing Drug Discovery. *Chem. Eng. News*, **2017**, *95* (25), 28–33.

[101] Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat. Rev. Drug Discov.*, **2016**, *16* (2), 131–147. https://doi.org/10.1038/nrd.2016.213.

[102] Decurtins, W.; Wichert, M.; Franzini, R. M.; Buller, F.; Stravs, M. A.; Zhang, Y.; Neri, D.; Scheuermann, J. Automated Screening for Small Organic Ligands Using DNA-Encoded Chemical Libraries. *Nat. Protoc.*, **2016**, *11* (4), 764–780. https://doi.org/10.1038/nprot.2016.039.

[103] Shi, Y.; Wu, Y. R.; Yu, J. Q.; Zhang, W. N.; Zhuang, C. L. DNA-Encoded Libraries (DELs): A Review of on-DNA Chemistries and Their Output. *RSC Adv.*, **2021**, *11* (4), 2359–2376. https://doi.org/10.1039/D0RA09889B.

[104] Castañón, J.; Román, J. P.; Jessop, T. C.; de Blas, J.; Haro, R. Design and Development of a Technology Platform for DNA-Encoded Library Production and Affinity Selection. *SLAS Discov.*, **2018**, *23* (5), 387–396. https://doi.org/10.1177/2472555217752091.

[105] Satz, A. L.; Kuai, L.; Peng, X. Selections and Screenings of DNA-Encoded Chemical Libraries against Enzyme and Cellular Targets. *Bioorg. Med. Chem. Lett.*, **2021**, *39*, 127851. https://doi.org/10.1016/J.BMCL.2021.127851.

[106] Dreiman, G. H. S.; Bictash, M.; Fish, P. V.; Griffin, L.; Svensson, F. Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding. *SLAS Discov.*, **2021**, *26* (2), 257–262. https://doi.org/10.1177/2472555220949495.

[107] 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* (5), 1255–1264. https://doi.org/10.1021/ACSCHEMBIO.6B00029.

**** An in-depth analysis on the methodological aspects of iterative HTS screening.**

[108] Cortés-Ciriano, I.; Firth, N. C.; Bender, A.; Watson, O. Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. *J. Chem. Inf. Model.*, **2018**, *58* (9), 2000–2014. https://doi.org/10.1021/ACS.JCIM.8B00376.

[109] Tidten-Luksch, N.; Grimaldi, R.; Torrie, L. S.; Frearson, J. A.; Hunter, W. N.; Brenk, R. IspE Inhibitors Identified by a Combination of In Silico and In Vitro High-Throughput Screening. *PLoS One*, **2012**, *7* (4), e35792. https://doi.org/10.1371/JOURNAL.PONE.0035792.

[110] 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* (13), 4891–4905. https://doi.org/10.1021/JM100488W.

[111] Jiang, P.; Zhang, D. Maternal Embryonic Leucine Zipper Kinase (MELK): A Novel Regulator in Cell Cycle Control, Embryonic Development, and Cancer. *Int. J. Mol. Sci. 2013, Vol. 14, Pages 21551-21560*, **2013**, *14* (11), 21551–21560. https://doi.org/10.3390/IJMS141121551.

[112] Liao, Z.; Gu, L.; Vergalli, J.; Mariani, S. A.; De Dominici, M.; Lokareddy, R. K.; Dagvadorj, A.; Purushottamachar, P.; McCue, P. A.; Trabulsi, E.; et al. Structure-Based Screen Identifies a Potent Small-Molecule Inhibitor of Stat5a/b with Therapeutic Potential for Prostate Cancer and Chronic Myeloid Leukemia. *Mol. Cancer Ther.*, **2015**, *14* (8), 1777–1794. https://doi.org/10.1158/1535-7163.MCT-14-0883.

[113] Wingelhofer, B.; Maurer, B.; Heyes, E. C.; Cumaraswamy, A. A.; Berger-Becvar, A.; de Araujo, E. D.; Orlova, A.; Freund, P.; Ruge, F.; Park, J.; et al. Pharmacologic Inhibition of STAT5 in Acute Myeloid Leukemia. *Leukemia*, **2018**, *32* (5), 1135–1146. https://doi.org/10.1038/s41375-017-0005-9.

[114] Orlova, A.; Wagner, C.; Araujo, E. D. de; Bajusz, D.; Neubauer, H. A.; Herling, M.; Gunning, P. T.; Keserű, G. M.; Moriggl, R. Direct Targeting Options for STAT3 and STAT5 in Cancer. *Cancers (Basel).*, **2019**, *11* (12), 1930.

https://doi.org/10.3390/CANCERS11121930.

[115] de Araujo, E. D.; Orlova, A.; Neubauer, H. A.; Bajusz, D.; Seo, H.-S.; Dhe-Paganon, S.; Keserű, G. M.; Moriggl, R.; Gunning, P. T. Structural Implications of STAT3 and STAT5 SH2 Domain Mutations. *Cancers (Basel).*, **2019**, *11* (11), 1757. https://doi.org/https://doi.org/10.3390/cancers11111757.

[116] Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science*, **2002**, *298* (5600), 1912–1934. https://doi.org/10.1126/science.1075762.

[117] Bajusz, D.; Ferenczy, G. G.; Keserű, G. M. Structure-Based Virtual Screening Approaches in Kinase-Directed Drug Discovery. *Curr. Top. Med. Chem.*, **2017**, *17* (20), 2235–2259. https://doi.org/10.2174/1568026617666170224121313.

[118] Cohen, P. Protein Kinases - the Major Drug Targets of the Twenty-First Century? *Nat Rev Drug Discov*, **2002**, *1* (4), 309–315.

[119] Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacol. Res.*, **2016**, *103*, 26–48. https://doi.org/10.1016/j.phrs.2015.10.021.

[120] Singh, N.; Sun, H.; Chaudhury, S.; AbdulHameed, M. D. M.; Wallqvist, A.; Tawa, G. A Physicochemical Descriptor-Based Scoring Scheme for Effective and Rapid Filtering of Kinase-like Chemical Space. *J. Cheminform.*, **2012**, *4*, 4.

[121] Yang, Y.; Zhang, Y.; Hua, Y.; Chen, X.; Fan, Y.; Wang, Y.; Liang, L.; Deng, C.; Lu, T.; Chen, Y.; et al. In Silico Design and Analysis of a Kinase-Focused Combinatorial Library Considering Diversity and Quality. *J. Chem. Inf. Model.*, **2020**, *60* (1), 92–107. https://doi.org/10.1021/ACS.JCIM.9B00841.

[122] Bajusz, D.; Ferenczy, G. G.; Keserű, G. M. Ensemble Docking-Based Virtual Screening Yields Novel Spirocyclic JAK1 Inhibitors. *J. Mol. Graph. Model.*, **2016**, *70*, 275–283. https://doi.org/10.1016/j.jmgm.2016.10.014.

[123] Bajusz, D.; Ferenczy, G. G.; Keserű, G. M. Discovery of Subtype Selective Janus Kinase (JAK) Inhibitors by Structure-Based Virtual Screening. *J. Chem. Inf. Model.*, **2016**, *56* (1), 234–247. https://doi.org/10.1021/acs.jcim.5b00634.

[124] Puhl-Rubio, A. C.; Stashko, M. A.; Wang, H.; Hardy, P. B.; Tyagi, V.; Li, B.; Wang, X.; Kireev, D.; Jessen, H. J.; Frye, S. V.; et al. Use of Protein Kinase Focused Compound Libraries for the Discovery of New Inositol Phosphate Kinase Inhibitors. *SLAS Discov.*, **2018**, *23* (9), 982. https://doi.org/10.1177/2472555218775323.

[125] Jacoby, E.; Wroblowski, B.; Buyck, C.; Neefs, J. M.; Meyer, C.; Cummings, M. D.; van Vlijmen, H. Protocols for the Design of Kinase-Focused Compound Libraries. *Mol. Inform.*, **2018**, *37* (5), 1700119. https://doi.org/10.1002/MINF.201700119.

[126] Natarajan, K.; Müller-Klieser, D.; Rubner, S.; Berg, T. Stafia-1: A STAT5a-Selective Inhibitor Developed via Docking-Based Screening of in Silico O-Phosphorylated Fragments. *Chem. – A Eur. J.*, **2020**, *26* (1), 148–154. https://doi.org/10.1002/CHEM.201904147.

[127] Wollenhaupt, J.; Metz, A.; Barthel, T.; Lima, G. M. A.; Heine, A.; Mueller, U.; Klebe, G.; Weiss, M. S. F2X-Universal and F2X-Entry: Structurally Diverse Compound Libraries for Crystallographic Fragment Screening. *Structure*, **2020**, *28* (6), 694-706.e5. https://doi.org/10.1016/j.str.2020.04.019.

[128] Skucha, A.; Ebner, J.; Schmöllerl, J.; Roth, M.; Eder, T.; César-Razquin, A.; Stukalov,

A.; Vittori, S.; Muhar, M.; Lu, B.; et al. MLL-Fusion-Driven Leukemia Requires SETD2 to Safeguard Genomic Integrity. *Nat. Commun.*, **2018**, *9* (1), 1983. https://doi.org/10.1038/s41467-018-04329-y.

[129] Skucha, A.; Ebner, J.; Grebien, F. Roles of SETD2 in Leukemia—Transcription, DNA-Damage, and Beyond. *Int. J. Mol. Sci.*, **2019**, *20* (5), 1029. https://doi.org/10.3390/ijms20051029.

[130] Castillo-González, D.; Mergny, J. L.; De Rache, A.; Pérez-Machado, G.; Cabrera-Pérez, M. A.; Nicolotti, O.; Introcaso, A.; Mangiatordi, G. F.; Guédin, A.; Bourdoncle, A.; et al. Harmonization of QSAR Best Practices and Molecular Docking Provides an Efficient Virtual Screening Tool for Discovering New G-Quadruplex Ligands. *J. Chem. Inf. Model.*, **2015**, *55* (10), 2094–2110. https://doi.org/10.1021/ACS.JCIM.5B00415/ASSET/IMAGES/LARGE/CI-2015-004158_0005.JPEG.

[131] Jang, W. D.; Jeon, S.; Kim, S.; Lee, S. Y. Drugs Repurposed for COVID-19 by Virtual Screening of 6,218 Drugs and Cell-Based Assay. *Proc. Natl. Acad. Sci. U. S. A.*, **2021**, *118* (30), e2024302118. https://doi.org/10.1073/PNAS.2024302118/SUPPL_FILE/PNAS.2024302118.SAPP.PDF.

[132] Zhang, M.; Zhu, W.; Li, Y. Discovery of Novel Inhibitors of Signal Transducer and Activator of Transcription 3 (STAT3) Signaling Pathway by Virtual Screening. *Eur. J. Med. Chem.*, **2013**, *62*, 301–310. https://doi.org/10.1016/j.ejmech.2013.01.009.

[133] Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Mol. Des.*, **2001**, *15* (5), 411–428.

[134] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.*, **2004**, *47* (7), 1739–1749.

[135] Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.*, **2004**, *47* (7), 1750–1759.

[136] Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.*, **2006**, *49* (21), 6177–6196. https://doi.org/10.1021/JM051256O.

[137] Miyao, T.; Funatsu, K. Iterative Screening Methods for Identification of Chemical Compounds with Specific Values of Various Properties. *J. Chem. Inf. Model.*, **2019**, *59* (6), 2626–2641. https://doi.org/10.1021/ACS.JCIM.9B00093.

[138] Szilágyi, K.; Flachner, B.; Hajdú, I.; Szaszkó, M.; Dobi, K.; Lőrincz, Z.; Cseh, S.; Dormán, G. Rapid Identification of Potential Drug Candidates from Multi-Million Compounds' Repositories. Combination of 2D Similarity Search with 3D Ligand/Structure Based Methods and In Vitro Screening. *Molecules*, **2021**, *26* (18), 5593. https://doi.org/10.3390/MOLECULES26185593.

[139] Szilágyi, K.; Hajdú, I.; Flachner, B.; Lőrincz, Z.; Balczer, J.; Gál, P.; Závodszky, P.; Pirli, C.; Balogh, B.; Mándity, I. M.; et al. Design and Selection of Novel C1s

Inhibitors by In Silico and In Vitro Approaches. *Molecules*, **2019**, *24* (20), 3641. https://doi.org/10.3390/MOLECULES24203641.

[140] Yan, X. C.; Sanders, J. M.; Gao, Y. D.; Tudor, M.; Haidle, A. M.; Klein, D. J.; Converso, A.; Lesburg, C. A.; Zang, Y.; Wood, H. B. Augmenting Hit Identification by Virtual Screening Techniques in Small Molecule Drug Discovery. *J. Chem. Inf. Model.*, **2020**, *60* (9), 4144–4152. https://doi.org/10.1021/ACS.JCIM.0C00113.

**** A perspective with three case studies on integrated VS and experimental screening.**

[141] Kutchukian, P. S.; Warren, L.; Magliaro, B. C.; Amoss, A.; Cassaday, J. A.; O'Donnell, G.; Squadroni, B.; Zuck, P.; Pascarella, D.; Culberson, J. C.; et al. Iterative Focused Screening with Biological Fingerprints Identifies Selective Asc-1 Inhibitors Distinct from Traditional High Throughput Screening. *ACS Chem. Biol.*, **2017**, *12* (2), 519–527. https://doi.org/10.1021/ACSCHEMBIO.6B00913.