Computational advances in combating colloidal aggregation in drug discovery

Daniel Reker,[1,*] Gonçalo J. L. Bernardes,[2,3] Tiago Rodrigues[3,*]

[1] Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[2] Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, UK.
[3] Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Av. Prof. Egas Moniz 1649-028 Lisboa, Portugal.

*Correspondence:
D.R.: reker@mit.edu
T.R.: tiago.rodrigues@medicina.lisboa.pt; tiago.rodrigues@alumni.ethz.ch

Competing interests:
The authors of this article declare that Vikki Cantrill is employed by Gonçalo Bernardes as a Research Coordinator at the University of Cambridge. Vikki Cantrill was not involved in the preparation, writing or editing of this article, but is married to Stuart Cantrill, who is the Chief Editor of Nature Chemistry.

Author contributions:
D.R. and T.R. contributed equally; conceived the study, performed literature and data analyses. D.R., G.J.L.B. and T.R. wrote the manuscript. Request for materials should be addressed to either D.R. or T.R. All authors approved the submitted version of the manuscript.

Abstract

Small molecule effectors are essential for drug discovery. Specific molecular recognition, reversible binding, and dose-dependency are usually key requirements to ensure utility of a novel chemical entity. However, artefactual frequent-hitter and assay interference compounds may divert lead optimization and screening programs towards attrition-prone chemical matter. Colloidal aggregates are the prime source of false positive readouts, either through protein sequestration or protein-scaffold mimicry. Nevertheless, assessment of colloidal aggregation remains somewhat overlooked and under-appreciated. In this Review we discuss the impact of aggregation on drug discovery by analysing select examples from the literature and publicly-available datasets. We also examine and comment on technologies used to experimentally identify these potentially problematic entities. We focus on evidence-based computational filters and machine learning algorithms that may be swiftly deployed to flag chemical matter and mitigate the impact of aggregates in discovery programs, and highlight the tools that can be used to scrutinize libraries, identify, and eliminate these problematic compounds.

Introduction

Specific binding of a drug to a given target and modulation of its activity remains the hallmark of clinical medicine[1]. To that end, small molecule entities have a long history and a paramount role as chemical probes and drug target effectors. Target- as well as cell-based screens are now privileged means of testing large compound collections in a high throughput manner to fuel development pipelines. Naturally, quality chemical probes and drug leads are in high demand to interrogate biological events, and infer the relationship between drug target binding and downstream phenotype changes[2,3]. However, the usefulness of unpruned small molecule libraries is usually questionable and can lead to artefact readouts. As a consequence, it is now well known that a significant proportion of high-throughput screening (HTS) hits are indeed nuisances – "false positives" – that can divert research efforts towards attrition prone chemical matter[4,5]. False positives can be loosely defined as compounds that show up as hits in the primary HTS assay readout without *de facto* exerting the desired biological effect. Often, this only becomes visible when they fail to produce the desirable target activity in orthogonal, follow-up validations.

False positives have shown to be abundant in discovery chemistry, having the potential to disrupt not only basic research in chemical biology and medicinal chemistry but also applied programs aiming at bringing small molecules from bench to clinic[5]. Strikingly, when gathering data on false positives to better understand their nature, it was discovered that they are not target-specific, but can be identified as "hits" against multiple biological targets, target-families, and even while using different assay technologies[6]. This observation has propelled recent investigations on assay interference molecules and the awareness of the structural underpinnings for undesirable assay behaviour[7-9]. For example, careful screening library design together with a trained medicinal chemistry eye appear to be central to remove problematic compounds and, thereby, mitigate pre-clinical attrition[5,10,11].

In line with this recent awareness, nine editors-in-chief of journals published by the American Chemical Society wrote an editorial to alert the need of ascertaining the quality of the reported chemotype–biology association data[12]. As pinpointed by those recommendations, one of the major reasons for false positive readouts is the aggregation of poorly water-soluble compounds into colloidal, nano- and micro-scale particles. These Small Colloidally Aggregating Molecules (SCAMs) can produce false positives in a broad range of assay formats, following well-established physical chemistry underpinnings[13]. In spite of being able to explain and control such effects, the molecular frameworks that can lead to aggregation are currently not understood. As a rule of thumb, molecules with high lipophilicity and/or surface area are more likely to aggregate in aqueous solution, and although potentially useful, special care must be taken to balance aggregation *vs.* potency and derisk the lead development process. For example, this can be achieved by reducing the aromatic ring count, which can positively impact on the kinetic solubility of small molecules[14]. However,

drugs frequently contain aromatic rings and not all of them present a liability, such that there is an urgent need for a more detailed understanding of the molecular frameworks causing aggregation. In contrast to assay interference compounds[7,8,15,16], *e.g.* compounds that interfere directly with the assay readout for example *via* auto-fluorescence, the colloidal aggregation phenomenon has been substantially less considered by many chemical probe and drug designers as attested by fewer publications and fewer available counter-screen efforts.

To bring attention to this problem and enhance the general understanding of the matter, we devote the following sections of the Review to this subject. By critically discussing select examples and analysing available data, we highlight the impact of colloidal aggregation in drug discovery. In particular, we focus on machine learning methods that can be swiftly deployed for the identification and flagging of aggregating chemical matter. We hope this Review will enable researchers at the forefront of drug discovery and chemical biology to understand and apply such methods and thus further their discovery programs by avoiding nuisance compounds.

Colloidal aggregates as source of false positive readouts

SCAMs can inhibit enzymatic activity in different functional assays. It has been shown that they have the potential to elicit reproducible, yet confounding and irrelevant enzymatic responses, given the unspecific nature of the molecular recognition mechanisms[17,18]. Formation of colloidal aggregates, unlike auto-fluorescence and reactivity, is largely independent of both the assay technology and drug target, but instead driven by the chemical structure of the assayed molecule and the buffered medium conditions. The latter include the pH value, buffer composition and concentration, and the amount and identity of surfactant in the buffer[19,20]. In analogy to micelle formation, the critical aggregation concentration (CAC) governs the assembly of those aggregates and its proportion to monomeric small molecule forms[21]. As soon as the testing concentration is above the CAC, nucleation occurs and colloidal aggregates start forming. The exact mechanism of how such aggregates form and their exact shape is still a matter of active research, but is well grounded on the physical chemistry of the small molecules[22]. The most recent progress suggests that colloidal aggregates form as unstructured, filled spheres, which can be composed by as many as 10^8 small molecules, and present large polydispersity[21]. Many of these aggregates seem to continue growing and eventually precipitate after reaching a given size[22]. What is certain is that colloidal aggregation drastically changes the ability of the chemical matter to interact with other macromolecules, possibly due to the large surface area and hydrophobicity. From a molecular point of view, the generated particles are usually present in the mid-femtomolar to low picomolar concentration range, and bind up to 10,000 enzymes per aggregate. Because of the tight binding, these aggregates promote

partial and local protein denaturation, resulting in the abovementioned enzyme/protein inhibition in a time-dependent manner (Figure 1)[21-24]. Recently, crystallography experiments suggested a complementary mechanism of action that originates from much smaller aggregates comprising only 5–6 molecules. These can bind to the protein surface and displace protein monomers, thereby inhibiting functional protein activity. An x-ray crystal of a TNFα–aggregate complex clearly shows a non-symmetrical small molecule conglomerate replacing one of the three TNFα subunits of the *apo* form. The authors suggest that, like in specific molecular recognition, the aggregate induces a quaternary structure change without sequestration nor denaturation[25] (Figure 1). Given that an x-ray structure captures only one diffraction snapshot, which is not always representative of the protein and ligand conformations in solution, one may still question if this mimicry mechanism is valid in biochemical, target-based assays.

The deceiving effects of aggregates can also extend to cellular assay systems, where they can lead to less-than-expected activity due to reduced cell membrane diffusion, and hence reduced intracellular concentration[26]. Furthermore, aggregates have the potential to artefactually inhibit membrane-bound proteins, *e.g.* G-protein coupled receptors (GPCRs), as reported by Sassano *et al*.[27] It has also been recognized that aggregation might indeed not only impact drug screening, but aggregates can also be observed in biologically relevant environments. For example, the formation of colloidal aggregates in simulated intestinal fluid may lead to unpredictable pharmacokinetics, including absorption and distribution[28]. Altogether, the pernicious effects of SCAMs in chemical probe and lead discovery cripple both target- and cell-based assay readouts. Therefore, their identification at the early stages of molecular design is essential.

The impact of SCAMs on drug discovery projects

As discussed above, there are different mechanisms by which SCAMs can influence assay readouts and afford false positives. This indeed suggests that numerous research efforts aiming to specifically perturb the function of drug targets are potentially affected by such molecules. It remains to be studied how frequent these occur and understand if warnings regarding colloidal aggregators are merely anecdotal and do not require further attention. Unfortunately, it turns out that molecular signatures necessary to form colloidal aggregates, although not fully understood mechanistically, are empirically prominent among relevant chemical matter in drug discovery and early development efforts. Shoichet and co-workers have pioneered the high-throughput detection of SCAMs, having identified several thousand of them among natural products[29], drugs[30], “chemical probes”[27], and molecules from vendor libraries[31]. SCAMs are frequently found in unpruned HTS and

vendor compound libraries, from where such false positive “hits” have been all-too-often reported in troubling proportions[17,19].

Specifically, the formation of colloidal aggregates has been identified as a major source of false positive readouts amongst micromolar HTS “hits”[19]. It has been estimated that 5–20%[32] of screening compounds possess the chemical signature to form aggregates under typical HTS conditions. While this number might be considered low, it can indeed lead to a large amount of false positives among the usually small number of total acquired screening hits. For example, Feng *et al.* reported that a staggering 95% of HTS hits present a colloidal aggregation signature[33]. Likewise, in a HTS program aiming at the identification of starting points for optimization against the cysteine protease cruzain, 90% of the original hits were colloidal aggregate artefacts[17]. In another campaign against the same protease, an oxadiazole hit was expanded to a small, focused library for structure-activity relationships (SAR) studies[31]. Interpretable SAR was obtained only to realize later that several of the library members acted in divergent mechanisms, including colloidal aggregation. Surprisingly, those aggregates were responsible for multiple log units of apparently “sound” SAR and were only unveiled as nuisances from steep Hill slopes that suggested superstoichiometry in the biochemical assay[31]. While SAR is often sought after and may be considered a hallmark of chemical series tractability, this particular study definitely suggests otherwise. Indeed, it further highlights the need for hit validation and the role of multiple assays to probe for confounding bioactivity profiles. Fascinatingly, such high-rates of false positives in classical screening libraries are not necessarily eliminated through advanced pre-selection techniques. For example, a report on the discovery of new ligands for the orphan GPCR37L1 used a sophisticated computational method that predicted interactions between ligands and binding pocket for the selection of promising candidates for biochemical screening. In spite of this mechanistic filtering, a total of two out of five acquired hits were SCAMs[34]. These results are worrisome for the community given that large fractions of discovery efforts are harnessing chemical screening libraries similar to the ones in these studies, which are expected to contain a similar fraction of potential SCAMs. With this in mind, one might wonder whether SCAMs are similarly prevalent in other chemical classes relevant for drug discovery.

Natural product-based medicines have played an important role in therapeutics for several millennia. Given the biological pre-validation of their architectures[35-38] they have also been used as screening compounds against targets of interest. Recently, 14 natural products whose bioactivity has been interrogated in hundreds of scholarly reports were analysed for colloidal aggregation[29]. Eight of those molecules, *e.g.* physcion and curcumin (Figure 2a), formed colloidal aggregates at concentrations within their bioactivity range – typically above 10 μM. For example, physcion begins aggregating at concentrations as low as 0.32 μM. Such observation might render multiple reported bioactivities as irrelevant. Importantly, curcumin has been studied over 6000 times, amassing several million dollars spent on research[39,40]. However,

several of these studies do not account for the colloidal aggregation signature of curcumin at the utilized concentrations[39], raising serious doubts about the biological relevance and therefore clinical translatability of the *in vitro* generated associations[29]. Admittedly, since CAC values depend on the exact experimental layout, it would be unwise to refute all previous bioactivity reports of these natural products. Even for aggregators there are likely conditions, targets, and a concentration range where this mechanism is not relevant. However, given the limited awareness of colloidal aggregation and the limited usage of detergents in biological assays, it is probable that many of the reported activities are irrelevant.

Our own natural product research has included routine tests to rule out colloidal aggregation-derived artefacts. We have found that some complex entities, *e.g.* archazolid A[41] and (–)-englerin A[42], do form aggregates at double-digit micromolar concentrations. However, aggregation at those concentrations does not interfere with bioactivity readouts in the low micromolar or nanomolar range. Conversely, fragment-like entities may be less prone to such a behaviour at high concentrations[43], but exceptions do exist[44]. It is, however, important to stress that given the long-standing tradition of utilizing natural products as chemical probes and leads enables us also to estimate the rate at which such potential false-positives are reported. For example, quercetin and staurosporine aglycone (K-252c) have been used to study phosphoinositide 3-kinase and protein kinase C, respectively, for several years. Against their main targets, both molecules have displayed IC_{50} values of 2.5–4 μM[45,46], which are in the range of activities typically seen for SCAMs; a flag should therefore be raised. Indeed, these "probes" have been confirmed to aggregate at relevant biological concentrations through multiple assays[47]. In 2003, quercetin and K-252c had been referenced 2,878 and 363 times in PubChem, respectively[47]. Strikingly, their citation count has now considerably increased even further to >14,000 times for quercetin and >650 times for K-252c. Many of these reported bioactivities are potentially erroneous. The fact that weak bioactivities for such structures keep appearing may also support the limited awareness of the implications of colloidal aggregates in chemical biology and drug discovery, and the implied skewed biological data. Further analyses on quercetin bioactivities annotated in ChEMBL shows a strong bias towards low affinity values, which may exclude true promiscuous inhibition and a privileged structure in drug discovery. In fact, >65% of those values lie in a range of potential colloidal aggregation, and hence artefactual target inhibition (Figure 2b)[47]. Moreover, a recent report shows that the quercetin behaviour can be extended to its scaffold class – flavonoids – raising questions on their tractability for chemical probe and hit-to-lead campaigns[20].

As it turns out, SCAMs are not only prevalent in screening libraries but some of them can be found among approved drugs. In a seminal contribution, a screen of 50 approved drugs revealed that aggregation is present among high value chemical matter (14%)[30]. This suggests that extensive medicinal chemistry may not be enough to completely abrogate aggregation, and that optimizing SCAMs into drugs is

likely to include additional development challenges. From the selected drugs, clotrimazole (Figure 2a) aggregated at concentrations below 100 μM. While these effects are unlikely to affect the potent interaction of these drugs with their on-targets, aggregation may be considered as a "feature" of small molecules, given that even highly optimized chemical matter might still aggregate and create false positive results if tested at sufficiently high concentrations. Such false positive readouts might, however, sabotage drug repurposing efforts as well as the selective optimization of side effect activities campaigns[48].

A decade ago, Hopkins coined the term "network pharmacology" as the underlying basis of polypharmacology and billing it as a paradigm for future drug discovery[49,50]. It appears logical that SCAMs may compromise the accuracy of established pharmacology networks by adding false target links. In fact, analysing a list of validated SCAMs[51], one may estimate that 45-80% of bioactivities in ChEMBL for SCAMs might constitute false positive (*cf.* Box 1), which will have a dramatic impact on drug target networks by adding erroneous protein associations. Specifically, computation of pharmacology networks as a function of ligand commonalities allows straightforward conclusions after binning data into two groups – very low (p*Affinity* < 4) and very high (p*Affinity* ≥ 7) affinities. Very low affinities possibly represents false positives, considering that >100 μM hits are more likely to result from unspecific binding. While likely false ligand–target associations constitute the bulk of the extracted data, totalling 3,855 potentially erroneous activities (Figure 2c), SCAMs can still afford 1,985 cases of specific, potent target binding (~1% of all target annotations for known SCAMs, p*Affinity* ≥ 7, Figure 2d). The latter include tests against 297 different targets and a network comprising 1,672 binary interactions, which provides unexpected opportunities for further biological exploration. The analysis further highlights that aggregation is a contextual problematic, warranting target binding validation on a case-by-case basis to ascertain true target engagement. Altogether, the data supports that ligand-centric drug target relationships are indeed partly skewed due to colloidal aggregation and that careful validation of target binding/engagement is required. Admittedly, the bias will be largely introduced for lower potency compounds; hence the impact on medicinal chemistry optimization might be limited. Furthermore, the presented networks specifically focus on SCAMs, which only compromise a subset of the available bioactivity data. However, the sheer number of edges and convolution of the presented plots, in particular for proteins that have been targeted frequently, can introduce a noticeable bias in polypharmacology networks and hamper future drug discovery and biochemistry by incorrectly emphasizing irrelevant target relationships and misleading SAR.

Experimental SCAM detection methods

Some direct observations on the experimental results might justify suspicion on whether the determined interaction is real or a result of aggregation. Most importantly, steep dose-response curves (high Hill coefficients) provide initial suspicion for colloidal aggregation, as sudden nucleation just above the CAC can lead to the formation of aggregates with apparent potent inhibitory effects. Indeed, thousands of enzymes in a biochemical assay can potentially be sequestered by a single aggregate[24]. It is important to keep in mind that, in some cases, such high slopes are not pathological but a consequence of drug target dimerization[52], or cooperative binding to allosteric sites[53]. Moreover, "bell-shaped" concentration-response curves in cell viability assays have also been partly linked to colloidal aggregation, whereby the negligible diffusion of colloidal aggregates across intact cell membranes, and hence a reduction of intracellular drug concentration, is responsible for such a behavior[26,54].

Several biophysical and biochemical technologies can and should be routinely used to identify the aggregation-based inhibition mechanism by bioactive molecules, especially to scrutinize micromolar hits. A validation of the observed activity can be performed by slightly altering the original assay conditions; it is well established that re-screens in presence of small amounts of detergent (*e.g.* 0.01% Triton X-100 for biochemical assays or 0.025% Tween-80 as a gentler detergent for cell-based assays[55]) can discriminate false from true positive hits; as detergents will favour the monomeric form of molecules in solution, their apparent activity (*e.g.* percentage inhibition of control) will be lower when such detergents are present in assay buffers. Evaluating whether the observed binding and/or target engagement is reduced or abolished with increasing detergent concentrations allows a conclusion on an unspecific target engagement event[31,56,57]. Indeed, performing parallel screens with and without detergent and relying on model enzymes, such as beta lactamase, has become one state-of-the-art protocol to identify colloidal aggregates (Table 1). Alternatively, instead of adding detergents, the addition of bovine serum albumin to the assay medium has been utilized as a means of attenuating the inhibition of proteins/enzymes by SCAMs through a competing adsorption mechanism at the aggregate's surface[58]. More recently, Adachi and co-workers[59] have developed a DMSO-perturbed assay to identify SCAMs among natural products. In short, the addition of DMSO to the assay mixture leads to conformational changes in the protein, and the formation of two distinct protein populations – one productive (folded) and another that is non-productive (partly unfolded). In the presence of SCAMs, non-specific binding to the non-productive enzyme population occurs, leading to a decrease in the effective concentration of ligand in solution and, consequently, its inhibitory activity against the folded population[59]. For all approaches, it has to be confirmed that the additive (detergent, DMSO or serum albumin) does not interfere with the readout, *e.g.* by interacting with the investigated protein itself. Furthermore, the difference in assay activity needs to be statistically

quantified, to ensure that the observed changes are not simply induced through experimental variations. This approach represents the most direct readout of false positive determination by relying on the assay of interest and closely mimicking the original screening conditions.

Other orthogonal assay technologies have been developed and can be employed to investigate a compound's potential to form colloidal aggregates. Dynamic light scattering (DLS) is able to identify the presence of colloidal aggregates, determine the underlying critical aggregation concentration and assess the kinetic solubility of chemical matter of interest[31,32,60]. Though this method has become the gold standard to test for aggregation there remain major challenges associated with current DLS setups. Because the DLS mathematical modelling of diffracted laser light assumes a certain particle shape, its measurements have demonstrated inconsistent readings for several SCAMs whose aggregates adopt shapes other than the predefined[61]. Thus, DLS may be valuable for qualitative studies but shows limitations for studying the exact size of colloidal aggregates. It is also not always trivial to represent the exact original assay conditions for the DLS measurement, especially when co-factors and proteins are contained in the assay buffer and medium, which can contribute to a high scattering background that overshadows the aggregate signal[62]. Moreover, not all SCAMs form colloidal particles measurable through DLS, which can significantly increase the percentage of false negative aggregation events[32]. The use of transmission electron microscopy may partly mitigate limitations of DLS, allowing the measurement of aggregates in physiologically-relevant media. However, microscopy imaging can be time consuming[63]. Importantly, observing aggregates through DLS or microscopy does not necessarily imply unspecific inhibition and thus false positive readouts. Apparently some aggregates lead to immediate precipitation or do not bind enzymes and thereby will not show up as unspecific hits in biochemical screening assays. Their relevance as chemical probes is nevertheless questionable. As an easy-to-implement routine, colloidal aggregates can be actively precipitated by centrifugation before an assay is run. Using this method, one can swiftly conclude on the formation of aggregates if the effectiveness of the compound against the cell- or target-based assay is much higher before centrifugation than after. Using a simple and sensitive fluorescence-based method, Cai *et al.*[64] have also been able to identify SCAMs. The method explores the surface tension changes in solvents as a function of colloidal alterations in solution. Formation of aggregates in solution significantly affects the shape of the meniscus and the fluorescence intensity when detected by a top-read fluorescence plate reader. Importantly, the method is amenable to high-throughput detection of SCAMs. Interestingly, colloidal aggregates originating from small molecules also afford unusual ^{1}H NMR resonances (at lower fields) and/or unusual peak shapes (broader). This becomes particularly visible across an increasing concentration series. This method is, however, only applicable to small size aggregates where tumbling is compatible with NMR timescales and has a restricted detection limit[65]. Albeit the lower throughput compared to most of the technologies cited above, the widespread use of NMR

spectrometers in medicinal chemistry laboratories renders it an easily accessible means to screen for potentially pathological chemical matter[65,66].

A new class of aggregation detection methods rely on advanced biosensor technology. Surface plasmon resonance (SPR) is a label-free technology commonly used to determine ligand–target binding kinetics and affinity constants in early drug discovery. Careful analysis of the resulting sensorgrams not only provides crucial knowledge of the stoichiometry of binding, but may also reveal pathological binding events as reported by Giannetti *et al*.[67] Using SPR, it was shown that colloidal aggregates may interact reversibly only with some proteins, for which the response units are not stoichiometric. Furthermore, the interaction mechanism varies between proteins, suggesting that it may be difficult to identify a method for the systematic identification of SCAMs. Photonic crystal (PC) optical biosensors are another label-free biophysical technology that has been productively employed to detect SCAMs with competitive results to other technologies[61]. The PC sensors comprise a sub-wavelength periodic surface that reflects a narrow band of wavelengths when illuminated with a broadband collimated light source. Upon binding of aggregates, the sensor surface is changed, modulating the refractive index of the material[61]. Among the many advantages of the method, it is important to highlight its independence relative to aggregate shape, and amenability to high-throughput screening.

While the methods discussed above may be in general easy to implement and can provide robust answers on the aggregation propensity of the molecules under investigation, running them can become unpractical, as in several cases they require expensive/dedicated equipment that is inaccessible to the wider community or require additional experiments with varying throughput. Accurate analyses of the generated data can also be time-consuming, deterring researchers from routine tests. Hence, they provide tools for hit validation, although their utility as a parallel assay for screening library curation might not be feasible. This stems from the fact that the abovementioned assays are often not high-throughput compatible, fail to reproduce the assay conditions, or pose concerns regarding their accuracy. Nonetheless, these methods offer an invaluable means of generating high-quality training data for tailoring *in silico* SCAM detection tools and streamline their identification. It will be important to see whether the data acquired from different screening methods is compatible and whether a learning algorithm can successfully determine relevant patterns.

Automated prediction of SCAMs

Rule-based cheminformatic filters[7] in drug discovery have traditionally provided medicinal chemists with tools to scrutinize small molecules. More recently, artificial intelligence has seen successful applications in the deconvolution of molecular

targets of phenotypic hits[35,41,60] and design new chemical entities[68], among others. Considering that the publicly available positive (aggregating molecules) and negative (non-aggregating molecules) data not only spans a wide range of calculated log*P* and molecular weight values (Box 1), but also notable scaffold diversity (Figure 3), one can envisage that machine learning algorithms may play an important role for the expeditious identification of potentially liable chemical matter. Those cheminformatics and machine learning approaches applied to flag screening compounds are discussed henceforth.

Despite being known that SCAMs can populate natural product and synthetic chemical spaces, very few publications report or analyse the exact structures of such aggregators. To the best of our knowledge, there is no structured and publicly available resource containing chemical structures of SCAMs, with the exception of datasets released by Shoichet and co-workers (shoichetlab.compbio.ucsf.edu). Exploration of those datasets allows concluding that between 2003 and 2015 the number of known aggregating chemotypes has grown exponentially; however as a result of only a few studies, *i.e.* knowledge in this particular field does not appear to be sustained but rather evolving in discrete studies through the introduction of assay technologies with increasing throughput. In particular, no in-depth scaffold analysis and its evolution in the past two decades has been researched. When investigating scaffold diversity on the Shoichet laboratory data, one can realize there is impressive scaffold diversity with 58% of the molecules featuring unmatched Murcko scaffolds among their aggregating counterparts (Figure 3). The data suggests that care was taken by Shoichet and colleagues throughout the years to test chemically diverse compounds. The scaffold diversity of SCAMs also highlights that the phenomenon of aggregation is not restricted to specific chemotypes but plagues a wide range of different chemical structures. Strikingly, flavones (Figure 3, scaffold count = 31) appear to be particularly susceptible to aggregation possibly as consequence of a flat/hydrophobic structure, which was accused previously of representing a chemical feature that promotes aggregation[69]. This structural feature is even more likely to be encountered in synthetic screening libraries. Altogether, analysis of the publicly available datasets also confirms that SCAMs are reasonably ubiquitous. Computational prediction methods can help to eliminate them from screening libraries or at least caution drug hunters in follow-up studies. Indeed, *in silico* tools promise a rapid and efficient way to predict colloidal aggregation or other liabilities from compound structures and prioritize hits for validation assays.

There is a long tradition of deploying computational reasoning for filtering unwanted structures in screening libraries both in the industrial and the academic sectors. Major research efforts have been devoted to standardizing the prediction of false positives using computationally-accessible guideline processing. Around 20 years ago, to translate chemical intuition into rapid, reproducible, and computable units, seminal work by Hann[10], Rishton[5] and Murcko[70-72] defined certain potentially critical chemical substructures as taboo lists. These lists were carefully designed to capture

substructures in screening molecules that could afford intractable or unfavourable readouts. These lists include moieties that can either decompose, react with assay components[73] or do not have drug-like properties (Box 2) – similar to the original "rule of five" developed by Lipinski[74] but focusing on the presence of alarming structures instead of physicochemical rules hardwired into the "rule of five". Using any of such or other proprietary substructure lists for the curation of screening pools has now become a commonplace practice[75]. More recently, Hajduk and co-workers developed a method named "ALARM NMR" to specifically assess thiol reactivity[9]. In an orthogonal approach, instead of focusing on the mechanistic filtering of reactive groups, Baell and Holloway followed a more empirical route and identified substructures of compounds that appeared as "hits" in multiple AlphaScreen assays and hence could represent target-agnostic false positives – coining them as pan-assay interference compounds (PAINS)[7].

Although frequently applied to filter or caution researchers against certain substructures, these flagging lists are inappropriate to detect SCAMs. Indeed, we emphasize that those lists were designed for a different purpose. For example, the assay technology utilized to detect PAINS specifically included Tween-20 to mitigate the effect of aggregators[76]. Comparing the number of alerts triggered for a dataset of known SCAMs compared to a set of random molecules of similar size and weight distribution that is largely devoid of SCAMs, we observed practically no difference between the number of flagged compounds between those two sets (Figure 4 and Box 1)[51]. With the exception of ALARM NMR, all filters consistently fail at identifying SCAMs, *i.e.* only low numbers of known SCAMs are flagged, and the number of alerts do not differ from the numbers of flagged compounds in the random background (ChEMBL-derived) set. This also implies that the number of other liabilities among SCAMs are equally frequent compared to non-aggregators. Interestingly, many aggregators and non-aggregators are flagged by ALARM NMR with SCAMs showing a slightly higher numbers of alerts. This observation would need further validation, but can suggest that ALARM NMR implicitly encodes some of the features potentially correlated with aggregation, albeit not designed for that purpose; whether chemical reactivity is linked to colloidal aggregation remains unknown. Other hypotheses could involve potential reactive compounds being included in the aggregator dataset or, conversely, the ALARM NMR flags being partly trained on aggregation instead of reactive compounds. Nonetheless, the data reinforces that the aggregation phenomenon is ubiquitous, and that seemingly "clean" structures may still provide false positive readouts through a confounding mechanism such as aggregation. Even more importantly, this highlights that simple flagging lists were not designed to capture the complex phenomena of compound aggregation. More advanced *in silico* models might be necessary to reliably identify and combat these structures.

Machine learning models have been trained on large screening datasets to predict whether a compound is likely to occur as a hit in multiple screens. These might either

be true promiscuous binders or compounds that interfere with multiple assay technologies. These so-called “frequent hitter” compounds share features that can be computationally identified on a substructural level by applying more or less advanced mathematical models[6,77,78]. In the particular case of flagging undesired chemical matter, self-organizing maps[6] and Bayesian learning[79] (Box 3) have proven useful for the identification of “frequent hitters”. While in the first case data is clustered, preserving local neighbourhoods, through an algorithm that is inspired by artificial neural networks, in the latter case, probability modelling drives the data association in a typically faster manner. Both tools seem promising at identifying compounds that can hit in multiple HTS campaigns, but they are commonly not able to distinguish between false positives (artefacts) and privileged scaffolds (promiscuous compounds with specific effects on multiple targets). The Hit Dexter webserver can flag 65% of known SCAMs as “frequent hitters” by using extremely randomized decision trees trained on the chemical structures of compounds frequently studied in biochemical assays[80]. This performance is much better compared to the flagging lists discussed previously, but it is also unspecific as a majority of the non-aggregating compounds are flagged (inner circle). Given that Hit Dexter has shown good specificity in regards to not flagging non-promiscuous compounds[80], the result might suggest that the ChEMBL-derived compounds (Box 1), while largely lacking chemical signatures leading to aggregation, potentially exhibit other mechanisms of true or false promiscuity (Figure 4). This strongly advocates for the development and deployment of specific aggregation prediction models to delineate the exact mechanism of artefactual target modulation and alert the project team of different, context-dependent liabilities. Moreover, the Hit Dexter server does not flag 35% of known SCAMs, which shows that not all known aggregating molecules have been extensively employed in biochemical assays and likely are not (yet) perceived as “frequent hitters”. The fact that two-thirds of known SCAMs are captured by these methods emphasizes two points. Firstly, it suggests that accurate identification of SCAMs is possible, even though the discussed models were not designed to predict aggregation – indicating that specific aggregation models might potentially show better and more specific performance. Secondly, many SCAMs have already been tested repeatedly in multiple screening campaigns and potentially been conceived as true hits, which further advocates for the pernicious impact of SCAMs in drug discovery as well as for the need of a highly-accurate computational model to devalidate these “hits”.

Indeed, a handful of studies have aimed at developing advanced machine learning models specifically for the prediction of the aggregation likelihood of a query compound. As in any other classification applications, the training data must contain molecules belonging to both classes (aggregators *vs.* non-aggregators). We describe the methods here in chronological order, which coincidentally also corresponds to an increasing order of data requirements. A pioneering effort at the automated identification of SCAMs used as little as 111 chemical entities and 260 physicochemical properties that were recursively partitioned to afford a tree-like

model[30] (Box 3). Given the small training dataset, one may speculate that 260 descriptors could potentially be correlated or irrelevant to distinguish aggregators from non-aggregators in this data. Indeed, refinement of the recursive partitioning (RP) models through elimination of redundant and non-readily accessible descriptors afforded a model that was able to correctly classify 94% of the chemical entities. Not surprisingly, the Daylight Clog*P* descriptor afforded the first partition (*i.e.* represented the most important descriptor to separate aggregators from non-aggregators) with a cut off of 3.6. The result is intuitive and shows how simple machine learning algorithms can easily grasp data patterns even in small datasets. The presence or absence of ionisable groups and the extent of conjugation also decisively contributed to the accuracy of the classification model, highlighting that lipophilicity alone is not a sufficiently descriptive property to separate aggregators from non-aggregators (*cf*. Box 1); other features that capture the interaction-capabilities of the compounds further refine the predictive capabilities. Importantly, the RP model outperformed straightforward similarity searches, suggesting that aggregation is widespread in chemical space and not confined to certain clusters or chemotypes (*cf*. Figure 3)[30]. While such a modelling approach delivered a human-interpretable model that captures important aspects of the aggregation propensity of compounds, the limited training data used for modelling makes it an unlikely candidate to accurately flag large screening libraries with thousands of compounds that differ substantially from the 111 compounds used for training. Indeed, in a follow-up study, the predictive utility of the RP model to correctly classify aggregators and non-aggregators in a large screening was refuted[32]. With the advent of more sophisticated experimental screening technologies for aggregators, specifically the possibility to compare HTS results with and without surfactants as an indicator of unspecific binding effects or using high-throughput DLS (Table 1), more data and hence more complex computational models became amenable. Such models promise to be able to capture a broader range of chemical signatures that will lead to aggregation. For example, the investigation of 1,030 drug-like molecules not only allowed constructing a naïve Bayes classifier, but also the evolution of the above discussed RP model into a more sophisticated random forest (RF) model, which affords predictions through a majority class vote within an ensemble of decision trees (Box 3)[32]. By relying both on chemical substructural fingerprints as well as physicochemical property descriptors, the RF achieved a low misclassification rate of 11%. In another effort at identifying SCAMs from physicochemical properties derived from chemical structure, Chen and co-workers[81] developed a support vector machine – recursive feature elimination (SVM-RFE, Box 3) classification model from 1,319 aggregators and 128,325 hypothetical non-aggregators that was competitive with other classifiers[81]. The authors used 199 molecular descriptors for model building, including simple molecular properties, connectivity, shape, electro-topological, quantum chemical and geometrical property descriptors to identify diverse SCAMs at the Murcko scaffold level. Gratifyingly, the feature elimination analysis corroborated all previous reports highlighting the relevance of hydrophobicity, connectivity, electro-topological, and quantum mechanical descriptors for characterizing SCAMs[81]. Thus, based on the

agreement of several independent studies using orthogonal algorithms, it is reasonable to assume that said descriptors are robust starting points for future investigations.

Although initial reports on sparse datasets had previously refuted the identification of aggregators from chemical structure similarity alone[30], more recent reports based on larger datasets show that model-free approaches simply based on molecular similarity to the reference data hold great potential for aggregation inference[51]. The Aggregator Advisor (http://advisor.bkslab.org) uses substructural fingerprints to determine similarity of query entities to over 12,000 experimentally validated SCAMs. A likelihood of aggregation for the queried compound is calculated by triaging the molecular similarity to known SCAMs, the observed affinity range, as well as the calculated log*P* of the query compound[51]. Considering there is no training step, the method is guaranteed to afford a reasonable amount of false positive/negative alerts, despite the remarkable diversity of physicochemical properties, except calculated log*P*, within the reference database. Furthermore, notwithstanding the high structural diversity of the molecules, as attested by a low average path-based fingerprint similarity, the reference database will always represent a minute percentage of chemical space, which will decisively contribute to said flagging inaccuracies.

To better understand how these different methods behave prospectively, we compared the three most promising methods, specifically the SVM-RFE[81], RF[32] classification models and the publicly-available Aggregator Advisor[51]. The use of all three methods for flagging of novel compounds from a ChEMBL subset of random molecules as aggregators (12,637 compounds, *cf.* Box 1) revealed a distinct behaviour between the different models on uncharted chemical structures. Not only did the algorithms differ in the molecules they would flag as aggregators, they also flagged vastly different numbers of compounds. Taking into account the number of predicted SCAMs for each model (Table 2), one may speculate that the RF approach is more conservative, with a potentially smaller fraction of false positive predictions compared to the remaining models but also the possibility of false negatives. This behaviour is also visible in the original publication through its low recall (60%) and high precision (83%) on the training data[32]. Therefore, the method is likely valuable when aggregators are to be identified with high confidence. However, this also means that many potential aggregators would not be flagged by this method. If flagging of chemical structures is imperative to generate exhaustive alerts and avoid any aggregators, it is arguably key to ensure a high recall. This would support the Aggregator Advisor or the SVM-RFE as models of choice. Indeed, the latter showed almost inverse performance statistics on its original training data compared to the RF model, with an impressive recall of 77% but a low precision of around 33%[81]. This indicates that relying on such a method could potentially eliminate thousands of promising structures through wrongly accusing them of aggregation – a behaviour that has brought classic flagging lists into disfavour recently (Box 2), and would likely also limit the deployment of such approaches in the pharmaceutical industry.

Understandably, despite this theoretical framework, only experimental characterization of the flagged chemical matter can provide a definitive answer.

We next wondered how different the chemical space landscape flagged by the different algorithms was. Both the Aggregator Advisor and SVM-RFE model predict more than 2,500 structures (>5% of the whole dataset) as aggregators, but have disparate classification vantage points, with only *ca.* 25% agreement (Figure 5a). The result is intriguing and raises questions on both the complementarity and ability to detect SCAMs by both methods. On the other hand, the RF model flags less than half the number of molecules compared to the other two approaches. An overlap of 68% with the SVM-RFE model supports a somewhat similar behaviour, which is not unexpected considering that they share the same training data set in the implementation. Given that the Aggregator Advisor is a model-free approach based on chemical similarity, this indicates that a total of 2,664 compounds in the prospective test set are chemically very similar to known aggregators. The RF and SVM-RFE models, on the other hand, generalize from physicochemical descriptors, thereby creating models that can predict SCAMs among novel chemotypes: 780 (70%) of all flagged compounds of the RF model and 1,718 (74%) for the SVM-RFE model are structures that are not highly similar to the training data. Conversely, 1,976 compounds are highly similar to the training data but not flagged by either the SVM or the RF models. This clearly highlights the differential characters of the different models, as well as the dissociation of the SVM and RF models from pure chemical similarity analysis. From a chemoinformatics perspective, one can also conclude that highly different model architectures with similar retrospective performances still contrast starkly in prospective behaviour and the character of flagged compounds. This can be easily visualized by using dimension reductions and projections of structural (dis)similarities between molecules identified as potential aggregators by the three different flagging methods (Aggregator Advisor, SVM-RFE, and RF models; Figure 5b). Even when relying on a completely independent definition of chemical similarity (*e.g.* MACCS keys) that is agnostic to the relationships implemented by the machine learning models, it becomes visible that the Aggregator Advisor largely flags highly-similar molecules with minute differences between them – visible through well-defined clusters among the flagged data (black dots in Figure 5b). Conversely, molecules identified by the SVM-RFE and RF models present a wider range of chemotypes as visible through larger dissimilarities between them, which can be attested to through the absence of clusters and a more widespread population of the chemical space. In summary, this result supports that SVM-RFE and RF models generalize learnt data patterns beyond simple chemical similarity, unlike the Aggregator Advisor that relies on chemical structure equivalence to alert on the aggregation propensity. Moreover, the chemical space projection also clearly highlights that the three methods behave differently and are able to provide distinct solutions to the challenge of identifying SCAMs with several molecules covering distinct areas, despite sharing the same or similar reference data.

Similar conclusions will be valid for alternative machine learning methods of potential interest for the colloidal aggregation problem, *e.g. k*-NN (Box 3), or the computationally simpler linear, and logistic regression methods. In any case, as model generalization is crucial for prospective utility, a fine balance between model complexity and accuracy must be found and ensure that the model is neither under- or overfit. The analysed machine learning models point towards the difficulty in identifying suitable descriptors as chemical markers for colloidal aggregation, and expose strengths but also limitations of the employed methods and algorithms. The significance of colloidal aggregation to mislead drug discovery and the limited number of current proposals to address this challenge highlights there is much room for improvement in the current machine learning applications. The publicly reported, successful applications of automated SCAM detection hint at established chemical descriptors coupled to advanced machine learning models being able to capture the aggregation propensity of a novel chemical structure, while algorithmic and performance shortcomings serve as important guideposts for future research in this most relevant area of drug discovery.

We must re-emphasize that, like for other substructural filters, the output of machine learning models of colloidal aggregation prediction should be interpreted with healthy scepticism and solely as an alert. As discussed above, all reported models still remain too crude to capture the concentration dependence of the colloidal aggregate formation *e.g.* via CACs[32]. The prediction of CAC values would require the employment of regression machine learning methods, which typically need even larger datasets for acceptable accuracy. The lack of this data and therefore requirement for extensive experimental profiling of CAC values to include in the training data can, at least in part, explain why such a regression model has never been reported before. Another shortcoming is the inability to distinguish different assay conditions, *i.e.* with diverging pH or buffer conditions the aggregation behaviour might not be relevant anymore for a compound of interest. Inclusion of the CAC value as well as the specific assay conditions might eventually provide the drug discovery community with advanced computational models to effectively gauge the relevance of aggregation propensity in a biological context. Such models and data will also improve our understanding of this phenomenon and provide useful guidance for the design of future HTS campaigns as well as the analysis of already published data. From a biological perspective, some enzymes might be particularly vulnerable to SCAMs while others might be less vulnerable to such effects[22]. Bootstrapping data from different campaigns will eventually enable us to develop a thorough understanding of when caution is advised.

Outlook

Modulation of drug targets in chemical biology and molecular medicine programs relies on the discovery and validation of new chemical entities that specifically bind the protein of interest. Evidence from literature, data analysis, and personal experience shows that the tortuous path to the successful discovery of disease-relevant chemical matter is often plagued by artefactual, false positive readouts. As discussed, colloidal aggregation can account for up to 95% of false positives in HTS campaigns, which renders it by far the most prevalent cause of erroneous ligand–target associations and yet, one of the least studied and tested for. We have here discussed select examples where discovery campaigns were misled by this effect and have estimated that up to 4% of screening data points are potentially compromised. This suggests that a staggering 80% of bioactivities for SCAMs might be false (Box 1). Overall, this alarming picture can compromise future drug discovery campaigns by inflating the selection of such seemingly promising structures. Tightly connected is the overlook of genuinely active compounds because of aggregation at sufficiently high micromolar concentrations, *i.e.* false negatives – this could be case for an undefined number of molecules. Both realizations advocate for the urgency of improving our understanding of the underpinnings of colloidal aggregation for informed decision-making.

To raise awareness on the impact of colloidal aggregates in discovery chemistry and provide tools to combat their development, we have here compared screening methodologies for the identification of SCAMs that may be valuable to the drug development community. A variety of methods exist with different complexities and advantages (Table 1). Their appropriate use will arm researchers with powerful tools to validate their screening hits. However, given the additional cost associated with counter-screens, emerging machine learning tools have a promising future for the automated aggregation prediction, to enable data curation, screening library filtering, and to study the aggregation behaviour further. Only few examples have been published on the automated detection of SCAMs (Table 2). While their performances are already promising and indicate that SCAM detection is indeed a computationally accessible challenge, their distinct vantage points and prospective behaviour indicate clear opportunities for improvement. Namely the gross under- or overestimation of the number of affected compounds or the strong dependence on high chemical similarity have important implications on their ability to predict the aggregation propensity of new chemical entities[82]. Therefore, it will be of utmost importance to obtain a thorough understanding of the behaviour and limitations of current methods, so that informed experts can apply and interpret their results and make decisions that suit the needs of their projects. We have shown here that available methods, albeit far from being perfect, already provide a range of different capabilities that will enable scientists to deploy these approaches with the desired outcome. Undeniably, with increasing data on aggregation and advances in predictive model architectures, the machine learning tools will become more

accurate at flagging liable chemical matter and overcome the current shortcomings of automated SCAM detection. In particular, with the advent of deep learning[83] and active learning[84] (Box 3) in medicinal chemistry and the life sciences in general, we expect the rapid development of more sophisticated models trained on large, high-quality data to emerge in the near future. Such optimized methods will provide invaluable *in silico* screening methods that act fast and are reliable to combat SCAMs by flagging liable structures for exclusion, re-screening, and validation. The identification of molecules forming colloidal aggregates does not always imply that they should be blindly discarded from further investigations. The bioactivity and the CAC should be evaluated *vis-à-vis* as a step towards proper validation of drug target binding and engagement. As discussed, drug target binding at low nanomolar concentrations is unlikely to be affected by aggregates if the CAC is orders of magnitude higher. In particular, given that many approved, life-changing drugs show aggregation at low concentrations indicates that such compounds can still act as potent therapeutics and their naïve elimination would dismiss potentially invaluable molecular material.

To promote well-informed decisions on this matter, we foresee that both human-controlled and "driverless" artificial intelligence will increasingly play a pivotal role in the setup and monitoring of drug discovery pipelines. Harnessing multiple predictive algorithms to anticipate crucial properties such as bioactivities, reactive accessibility, solubility, but also colloidal aggregation propensity will assist researchers in the efficient navigation of complex data pattern landscapes and streamline future drug discovery.

Box 1. Datasets

A random ChEMBL dataset[85] (green) of equal size (12,637 compounds) to the Shoichet database of validated SCAMs[51] (orange) was generated to mimic the molecular weight distribution of the latter. The obtained library (Figure above) includes 119 known SCAMs and 2,664 entities that are structurally very similar to known aggregators. This observation is fully in line with previous findings by Shoichet *et al.* on the expected frequency of SCAMs in publicly available databases[51]. Interestingly, although not selected for, a similar calculated log*P* (clog*P*) distribution emerged (Figure above), which suggests that molecular weight and predicted lipophilicity are correlated features in this particular library. Moreover, while there is a slight bias towards higher clog*P* values for the known aggregators, the distributions are strongly overlapping and therefore raise doubts about the utility of simple clog*P* measures to estimate the aggregation propensity of a molecular structure. Strikingly, disparate affinity profiles [p*Affinity* = –log(XC_{50} or $K_{D/i}$)] between the two subsets can be observed (Figure above). By distinguishing high from low potency values (cut off p*Affinity* = 6, *i.e.* 1 μM), some interesting trends became visible. Both sets showed an equal number of high affinity interactions, suggesting that the aggregation potential occurring at high ligand concentrations does seemingly not sabotage the potential of a compound to form specific, potent interactions with proteins. This is in line with observations that potent, approved drugs still aggregate at high concentrations. However, when looking at low affinity interactions, a stark difference was apparent. The average number of reported low affinity interactions was more than two times higher for aggregators compared to non-aggregators, hinting at 45% to 80% of the reported bioactivities for these structures being potentially false – depending on whether only assuming the additional low affinity interactions to be false or all low affinity interactions. Apart from economic implications, these wrong associations impact the development process of medications. Such potential false positives can dramatically perturb structure-activity relationships, misguide *in silico* method development and disrupt future drug discovery.

Box 2. Assay interference compounds

Especially since the introduction of PAINS, substructure filters have (re-)gained immense attention and are now more commonly applied. More than 1400 citations (as of December 2018; GoogleScholar) of the original PAINS paper attest to their frequent application and their utility for researchers to filter screening libraries to avoid artefacts. Indeed, leading journals, such as those published by the American Chemical Society, require authors to consider PAINS when publishing new bioactivities. However, as with any other technology, PAINS filtering is not a panacea[12] and they recently have been more thoroughly analysed and criticised[86,87]. It is now well established that substructure filters should not be applied blindly[76];

multiple retrospective analyses have disclosed that many flagged compounds will remain inactive in a broad range of assays[88,89]. This has been rationalized with the fact that some PAINS present features that are not necessarily associated to frequent-hitter behaviour[76,90]. Thus, not all PAINS represent liabilities irrespective of the molecular and screening context[88]. Raising further doubt about the utility of simple flagging lists, molecular design may provide a means to tailor molecules containing problematic moieties, and ultimately render them suitable for the clinics[91]. Indeed, there are multiple examples of "ugly" chemical matter with proven utility[92], including several FDA-approved natural products and derivatives, *e.g.* atovaquone, mitomycin, doxorubicin[91]. According to blindly following the abovementioned filters, these drugs would have been discarded from development pipelines, which would have significantly impacted the life of patients relying on these medications[7,8,15]. This should not be considered a refutation of the utility of PAINS, but a word of precaution regarding how the filters can be wrongly applied. They are useful formalizations of automated, focused chemical intuition[90], but blindly removing chemical matter from follow-up studies without sound context-based criteria could drastically influence future drug discovery. Instead, these rules should rather serve as guidelines to prioritize chemical matter and as a flagging system for downstream in-depth hit validation. Experimental validation of target engagement[93,94] is always warranted – a practice commonplace in the pharmaceutical industry. The figure depicts a taboo list (outer red circle) including examples of overlapping or unique motifs from the Hann[10], Rishton[5], "rapid elimination of swill" (REOS)[70,72] and "pan-assay interference compounds" (PAINS)[7,15] filters. Examples of FDA-approved drugs containing potentially problematic moieties are shown for comparison (inner green circle).

Box 3. A primer on artificial intelligence and machine learning

With the increasing amount of available data in drug discovery, new, fast and dependable ways of structuring it and exploring patterns became both needed and relevant. Data mining, machine learning, and artificial intelligence (AI) are overlapping technologies for automated data analyses in a holistic/integrated manner without the caveats of biased human reductionism[95]. At its core, AI software reacts to a previously "unknown" external stimulus (data features or descriptors) to provide an output (prediction) through a defined effector function (algorithm). AI leverages hundreds to millions of data points encoded through descriptors and use one of a dozen different mathematical approaches to recognize patterns or determine data correlations[95].

Machine learning methods include algorithms for grouping data into sets (clustering), to facilitate data interpretation and visualization (dimensionality reduction), to attribute a category to data objects (classification) and to predict numerical values (regression). In drug discovery, machine learning has seen applicability in the

prediction of solubility[96] or biological activity of drugs[97], the structure of proteins from their genetic sequence[98], or automated processing of medical imaging[99]. More recently, deep learning has emerged as a subfield of AI[100] that harnesses several hidden layers to unravel complex data. This method is especially data-avid and requires higher-than-usual computing power, which can potentially prohibit its use, despite notorious applications[83,101]. Conversely, active learning (or "selective sampling") focuses on building competent predictive models through selection of data points for testing and feeding them back into the model with the aim of modifying it in iterative fashion. It thus leverages a scarce amount of data, while using well-established algorithms to efficiently design experiments and add relevant knowledge[84,102,103].

Self-Organizing Maps (SOMs) or Kohonen networks[104] were developed as a neural network-inspired heuristic (with input/output neurons and their weight parameters) to reduce dimensionality and cluster data. As an unsupervised clustering method there is no target output, *i.e.* learning considers only the structure of the data but not the classes. Training evolves positions of cluster centroids ("neurons") until stabilization or a user-defined number of iterations. SOMs include a neighbourhood structure between the clusters and, within this neighbourhood network, small topological distances between clusters translate into small distances between cluster members in the original parameter space. Hence, SOMs function as a visualization technique by showing clusters with their neighbourhood relationship.

Naïve Bayes (NB) classifier is a supervised machine learning method that uses the Bayes theorem (assumption that the probability of any given event is correlated with a previously known variable) as workhorse to assess the degree of belief (the so-called posterior probability) in the association. The method assumes (naïve) independence between feature pairs to simplify joint probabilities $P(x,y)$ into products of independent probabilities $P(x)P(y)$. NB classifiers are highly versatile, robust and easy to use for the aggregation or other drug discovery problems by predicting the probability of a given class, *e.g.* aggregator/non-aggregator.

Recursive Partitioning (RP) or decision tree is a non-parametric and supervised method that can be used for classification and regression problems. It uses decision rules (split point) to separate the training data in an iterative (recursive) fashion, which leads to successively smaller groups of data that are more homogeneous in terms of the objective variable. From a tree model it is possible to extract the feature importance towards the prediction, *e.g.* by evaluating the order of splitting or calculate the associated impurity reduction of a variable. Decision trees do however tend to overfit. To mitigate this limitation, pruning of trees or ensemble methods have become increasingly popular.

Random Forests (RF) employ a combination (ensemble) of multiple tree predictors. Every tree is fitted using a randomly selected subset of the data and the available

data descriptors. Thereby, every tree learns to distinguish a certain aspect of the dataset, which contributes to high variance between the trees and mitigates their overfitting. When predicting new data, every tree will provide a prediction from their perspective, and the combination of all trees gives a consensus prediction (forest). Such a consensus can, for example, be the average predicted value (regression) or the majority class (classification). RFs are robust with respect to data noise, correlated features and descriptors with different magnitudes, which make them methods of choice for a vast majority of pattern recognition problems.

Extremely Randomized Trees (ERT) employ multiple decision tree predictors similar to RFs. The main difference lies in searching the most discriminative threshold for a certain variable. While RF searches all thresholds systematically, the ERT approach selects thresholds at random. The best thresholds are then used as splitting rules. This leads to a further de-correlation of the trees, which can improve the predictive ability of the ensemble.

Support Vector Machines (SVM)[105] are a method for classification or regression analysis of data projected into high (p) dimensional space. It identifies a so-called "hyperplane" in p–1 dimensional space that separates the training data, *e.g*. aggregators/non-aggregators. Multiple hyperplanes are thus possible with the optimal linear model presenting the maximum cumulative distance (margin) possible to the closest data points (support vectors). Since data might not be separable in the original parameter space, increasing the dimensionality to find separating planes is performed. Data separation can be achieved by implicitly increasing the dimensionality through the use of kernel functions. SVMs are sensitive to descriptor scaling, its performance is affected by the choice of parameters, can be slow to train, but have a lower risk of overfitting and work well on a wide range of problems.

Whereas RFs have an in-built feature selection process, the same is not true for SVMs[106], which can negatively impact the predictive performance. **Recursive Feature Elimination (RFE)** is a method to iteratively build smaller groups of features responsible for a given output. Pruning of the feature set is based on their relative weights (importance), with the first eliminated feature corresponding to that with smallest weight. The pruning procedure is repeated until a user-defined number of features is reached. Pre-processing the input data with a feature selection method may thus be not only desirable but also important. Alternative approaches for such pre-processing methods include **removing correlated features** or using **principle component analysis (PCA)** to reduce the dimensionality of the feature space.

***k*-Nearest Neighbour (*k*-NN)** is a supervised method using lazy learning (generalization is performed only once the algorithm is fed with a query) for classification or regression analysis of data by drawing conclusions from the k most similar reference data points. In both instances, the value of k is assigned by the user and can be weighted to reflect the distance between the query and the

reference data point. The **molecular similarity evaluation** can be considered a *k*-NN approach with *k* = 1 and using a distance threshold to ensure that the considered reference data point is sufficiently similar to be relevant for classification.

Irrespective of the employed classification or regression method, it is good practice to probe the scope and performance of the AI model. A ***k*-fold cross-validation** is commonly performed as a retrospective model evaluation approach to validate machine learning models. To that end, the dataset is partitioned into *k* equal-sized subsets, where one of those is held out for model validation, and the remaining *k*–1 subsamples are used to train the model to be tested. By performing a total of *k* evaluations, the whole data is used as a test set exactly once. A performance measure can then be calculated from the whole data, or on every individual evaluation and averaged. Important metrics to assess the utility of the classification models include:

- **Sensitivity or recall** is the fraction of positives identified as such. It is calculated as,

$$Sensitivity = \frac{TP}{TP + FN}$$

- **Specificity** is the fraction of negatives identified as such. It is calculated as,

$$Specificity = \frac{TN}{TN + FP}$$

- **Accuracy** is the fraction of correctly identified positives and negatives among the total population. It is calculated as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Matthews Correlation Coefficient (MCC)** gauges the quality of the model in a range from -1 to 1, where -1 denotes a completely wrong binary classifier and 1 a correct one. It stems from a 2×2 confusion matrix of true and false positives and negatives. It can be calculated as,

$$MCC = \frac{TP \times TN - FP \times FN}{((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{1/2}}$$

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively.

Regression methods may be best assessed through metrics such as:

- **Coefficient of determination (r^2)** can be determined between actual and predicted values, which reflects a better fit of the model to the training data through increasing values, up to a maximum of 1.

- **Mean absolute error (MAE)** measures the average error extent in a prediction set without considering its direction and assigning equal weights to the individual errors. It can be calculated as,

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

where n is the number of samples, y is a true value and $\hat{y}_i$ its paired predicted value.

Acknowledgments

D.R. is a Swiss National Science Foundation Fellow (Grants P2EZP3_168827 and P300P2_177833). G.J.L.B. is a Royal Society URF (UF110046 and URF\R\180019), an iFCT Investigator (IF/00624/2015), and the recipient of an ERC StG (TagIt, Grant Agreement 676832). T.R. and G.J.L.B. acknowledge Marie Sklodowska-Curie ITN Protein Conjugates (Grant Agreement 675007) for funding. T.R. is a Marie Curie Fellow (Grant Agreement 743640). T.R. acknowledges the H2020 (TWINN-2017 ACORN, Grant Agreement 807281) and POR Lisboa 2020/FEDER (02/SAICT/2017, Grant Agreement Lisboa-01-0145-FEDER-028333) for funding. The authors thank the comments and constructive criticism made by three anonymous reviewers.

References

1 Schurmann, M., Janning, P., Ziegler, S. & Waldmann, H. Small-molecule target engagement in cells. *Cell Chem. Biol.* **23**, 435-441 (2016).

2 Arrowsmith, C. H. *et al.* The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536-541 (2015).

3 Garbaccio, R. M. & Parmee, E. R. The impact of chemical probes in drug discovery: a pharmaceutical industry perspective. *Cell Chem. Biol.* **23**, 10-17 (2016).

4 Sink, R., Gobec, S., Pecar, S. & Zega, A. False positives in the early stages of drug discovery. *Curr. Med. Chem.* **17**, 4231-4255 (2010).

5 Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **2**, 382-384 (1997).

6 Roche, O. *et al.* Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **45**, 137-142 (2002).

7 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.* **53**, 2719-2740 (2010).

8 Devine, S. M. *et al.* Promiscuous 2-aminothiazoles (PrATs): a frequent hitting scaffold. *J. Med. Chem.* **58**, 1205-1214 (2015).

9 Huth, J. R. *et al.* ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* **127**, 217-224 (2005).

10 Hann, M. *et al.* Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **39**, 897-902 (1999).

11 Dahlin, J. L., Inglese, J. & Walters, M. A. Mitigating risk in academic preclinical drug discovery. *Nat. Rev. Drug Discov.* **14**, 279-294 (2015).

12 Aldrich, C. *et al.* The ecstasy and agony of assay interference compounds. *ACS Cent. Sci.* **3**, 143–147 (2017).

13 Ganesh, A. N., Donders, E. N., Shoichet, B. K. & Shoichet, M. S. Colloidal aggregation: From screening nuisance to formulation nuance. *Nano Today* **19**, 188-200 (2018).

14 Young, R. J., Green, D. V., Luscombe, C. N. & Hill, A. P. Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. *Drug Discovery Today* **16**, 822-830 (2011).

15 Baell, J. & Walters, M. A. Chemistry: Chemical con artists foil drug discovery. *Nature* **513**, 481-483 (2014).

16 Baell, J. B. Feeling Nature's PAINS: natural products, natural product drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **79**, 616-628 (2016).

17 McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45**, 1712-1722 (2002).

18 McGovern, S. L., Helfand, B. T., Feng, B. & Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **46**, 4265-4272 (2003).

19 Jadhav, A. *et al.* Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* **53**, 37-51 (2010).

20 Pohjala, L. & Tammela, P. Aggregating behavior of phenolic compounds--a source of false bioassay results? *Molecules* **17**, 10774-10790 (2012).

21 Coan, K. E. D. & Shoichet, B. K. Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.* **130**, 9606-9612 (2008).

22 Duan, D. *et al.* Internal structure and preferential protein binding of colloidal aggregates. *ACS Chem. Biol.* **12**, 282-290 (2017).

23 Coan, K. E. D., Maltby, D. A., Burlingame, A. L. & Shoichet, B. K. Promiscuous aggregate-based inhibitors promote enzyme unfolding. *J. Med. Chem.* **52**, 2067-2075 (2009).

24 Shoichet, B. K. Interpreting steep dose-response curves in early inhibitor discovery. *J. Med. Chem.* **49**, 7274-7277 (2006).

25 Blevitt, J. M. *et al.* Structural basis of small-molecule aggregate induced inhibition of a protein−protein interaction. *J. Med. Chem.* **60**, 3511-3517 (2017).

26 Owen, S. C. *et al.* Colloidal drug formulations can explain "bell-shaped" concentration-response curves. *ACS Chem. Biol.* **9**, 777-784 (2014).

27 Sassano, M. F., Doak, A. K., Roth, B. L. & Shoichet, B. K. Colloidal aggregation causes inhibition of G protein-coupled receptors. *J. Med. Chem.* **56**, 2406-2414 (2013).

28 Doak, A. K., Wille, H., Prusiner, S. B. & Shoichet, B. K. Colloid formation by drugs in simulated intestinal fluid. *J. Med. Chem.* **53**, 4259-4265 (2010).

29 Duan, D., Doak, A. K., Nedyalkova, L. & Shoichet, B. K. Colloidal aggregation and the in vitro activity of traditional chinese medicines. *ACS Chem. Biol.* **10**, 978-988 (2015).

30 Seidler, J., McGovern, S. L., Doman, T. N. & Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **46**, 4477-4486 (2003).

31 Ferreira, R. S. *et al.* Divergent modes of enzyme inhibition in a homologous structure-activity series. *J. Med. Chem.* **52**, 5005-5008 (2009).

32 Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K. & Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **1**, 146-148 (2005).

33 Feng, B. Y. *et al.* A high-throughput screen for aggregation-based inhinition in a large compound library. *J. Med. Chem.* **50**, 2385-2390 (2007).

34 Ngo, T. *et al.* Orphan receptor ligand discovery by pickpocketing pharmacological neighbors. *Nat. Chem. Biol.* **13**, 235-242 (2017).

35 Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531-541 (2016).

36 van Hattum, H. & Waldmann, H. Biology-oriented synthesis: harnessing the power of evolution. *J. Am. Chem. Soc.* **136**, 11853-11859 (2014).

37 Tannert, R. *et al.* Synthesis and structure-activity correlation of natural-product inspired cyclodepsipeptides stabilizing F-actin. *J. Am. Chem. Soc.* **132**, 3063-3077 (2010).

38 Takayama, H. *et al.* Discovery of inhibitors of the Wnt and Hedgehog signaling pathways through the catalytic enantioselective synthesis of an iridoid-inspired compound collection. *Angew. Chem. Int. Ed.* **52**, 12404-12408 (2013).

39 Nelson, K. M. *et al.* The essential medicinal chemistry of curcumin. *J. Med. Chem.* **60**, 1620–1637 (2017).

40 Baker, M. Deceptive curcumin offers cautionary tale for chemists. *Nature* **541**, 144-145 (2017).

41 Reker, D. *et al.* Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **6**, 1072-1078 (2014).

42 Rodrigues, T. *et al.* Unveiling (-)-englerin A as a modulator of L-type calcium channels. *Angew. Chem. Int. Ed.* **55**, 11077-11081 (2016).

43 Rodrigues, T., Reker, D., Kunze, J., Schneider, P. & Schneider, G. Revealing the macromolecular targets of fragment-Like natural products. *Angew. Chem. Int. Ed.* **54**, 10516-10520 (2015).

44 Rodrigues, T. *et al.* Machine intelligence decrypts β-lapachone as an allosteric 5-lipoxygenase inhibitor. *Chem. Sci.* **9**, 6899-8903 (2018).

45 Matter, W. F., Brown, R. F. & Vlahos, C. J. The inhibition of phosphatidylinositol 3-kinase by quercetin and analogs. *Biochem. Biophys. Res. Commun.* **186**, 624-631 (1992).

46 Fabre, S., Prudhomme, M. & Rapp, M. Protein kinase C inhibitors; structure-activity relationships in K252c-related compounds. *Bioorg. Med. Chem.* **1**, 193-196 (1993).

47 McGovern, S. L. & Shoichet, B. K. Kinase inhibitors: not just for kinases anymore. *J. Med. Chem.* **46**, 1478-1483 (2003).

48 Wermuth, C. G. Selective optimization of side activities: the SOSA approach. *Drug Discovery Today* **11**, 160-164 (2006).

49 Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682-690 (2008).

50 Hopkins, A. L. Network pharmacology. *Nat. Biotechnol.* **25**, 1110-1111 (2007).

51 Irwin, J. J. *et al.* An Aggregation Advisor for ligand discovery. *J. Med. Chem.* **58**, 7076-7087 (2015).

52 Mysinger, M. M. *et al.* Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5517-5522 (2012).

53 Prinz, H. Hill coefficients, dose-response curves and allosteric mechanisms. *J. Chem. Biol.* **3**, 37-44 (2010).

54 Owen, S. C., Doak, A. K., Wassam, P., Shoichet, M. S. & Shoichet, B. K. Colloidal aggregation affects the efficacy of anticancer drugs in cell culture. *ACS Chem. Biol.* **7**, 1429-1435 (2012).

55 Irwin, J. J. & Shoichet, B. K. Docking screens for novel ligands conferring new biology. *J. Med. Chem.* **59**, 4103-4120 (2016).

56 Ryan, A. J., Gray, N. M., Lowe, P. N. & Chung, C. W. Effect of detergent on "promiscuous" inhibitors. *J. Med. Chem.* **46**, 3448-3451 (2003).

57 Feng, B. Y. & Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **1**, 550-553 (2006).

58 Coan, K. E. & Shoichet, B. K. Stability and equilibria of promiscuous aggregates in high protein milieus. *Mol. Biosyst.* **3**, 208-213 (2007).

59 Tomohara, K., Ito, T., Onikata, S., Kato, A. & Adachi, I. Discovery of hyaluronidase inhibitors from natural products and their mechanistic characterization under DMSO-perturbed assay conditions. *Bioorg. Med. Chem. Lett.* **27**, 1620-1623 (2017).

60 Rodrigues, T. *et al.* Multidimensional de novo design reveals 5-HT2B receptor-selective ligands. *Angew. Chem. Int. Ed.* **54**, 1551-1555 (2015).

61 Chan, L. L. *et al.* A method for identifying small-molecule aggregators using photonic crystal biosensor microplates. *JALA* **14**, 348-359 (2009).

62 Rausch, K., Reuter, A., Fischer, K. & Schmidt, M. Evaluation of nanoparticle aggregation in human blood serum. *Biomacromolecules* **11**, 2836-2839 (2010).

63 Ganesh, A. N., McLaughlin, C. K., Duan, D., Shoichet, B. K. & Shoichet, M. S. A new spin on antibody-drug conjugates: Trastuzumab-Fulvestrant colloidal drug aggregates target HER2-positive cells. *ACS Appl. Mater. Interfaces* **9**, 12195-12202 (2017).

64 Lifeng, C. & Gochin, M. Colloidal aggregate detection by rapid fluorescence measurement of liquid surface curvature changes in multiwell plates. *J. Biomol. Screen.* **12**, 966-971 (2007).

65 LaPlante, S. R. *et al.* Compound aggregation in drug discovery: implementing a practical NMR assay for medicinal chemists. *J. Med. Chem.* **56**, 5142-5150 (2013).

66 Zega, A. NMR methods for identification of false positives in biochemical screens. *J. Med. Chem.* **60**, 9437-9447 (2017).

67 Giannetti, A. M., Koch, B. D. & Browner, M. F. Surface plasmon resonance based assay for the detection and characterization of promiscuous inhibitors. *J. Med. Chem.* **51**, 574-580 (2008).

68 Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).

69 Frenkel, Y. V. *et al.* Concentration and pH dependent aggregation of hydrophobic drug molecules and relevance to oral bioavailability. *J. Med. Chem.* **48**, 1974-1983 (2005).

70 Walters, W. P., Murcko, A. A. & Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **3**, 384-387 (1999).

71 Walters, W. P. & Namchuk, M. Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2**, 259-266 (2003).

72 Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening - an overview. *Drug Discovery Today* **3**, 160-178 (1998).

73 Olson, M. E. *et al.* Oxidative reactivities of 2-furylquinolines: ubiquitous scaffolds in common high-throughput screening libraries. *J. Med. Chem.* **58**, 7419-7430 (2015).

74 Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44**, 235-249 (2000).

75 Brenk, R. *et al.* Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **3**, 435-444 (2008).

76 Baell, J. B. & Nissink, J. W. M. Seven year itch: Pan-Assay Interference Compounds (PAINS) in 2017-utility and limitations. *ACS Chem. Biol.* **13**, 36-44 (2018).

77 Sadowski, J. & Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325-3329 (1998).

78 Schneider, P., Rothlisberger, M., Reker, D. & Schneider, G. Spotting and designing promiscuous ligands for drug discovery. *Chem. Commun.* **52**, 1135-1138 (2016).

79 Yang, J. J. *et al.* Badapple: promiscuity patterns from noisy evidence. *J. Cheminf.* **8**, 29 (2016).

80 Stork, C. *et al.* Hit Dexter: A machine-learning model for the prediction of frequent hitters. *ChemMedChem* **13**, 564-571 (2018).

81 Rao, H. *et al.* Identification of small molecule aggregators from large compound libraries by support vector machines. *J. Comput. Chem.* **31**, 752-763 (2010).

82 Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4067-4072 (2014).

83 Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* (2018).

84 Reker, D., Schneider, P. & Schneider, G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem. Sci.* **7**, 3919-3927 (2016).

85 Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945-D954 (2017).

86 Capuzzi, S. J., Muratov, E. N. & Tropsha, A. Phantom PAINS: problems with the utility of alerts for Pan-Assay INterference CompoundS. *J. Chem. Inf. Model.* **57**, 417-427 (2017).

87 Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **57**, 2640-2645 (2017).

88 Jasial, S., Hu, Y. & Bajorath, J. How frequently are Pan-Assay Interference Compounds active? Large-scale analysis of screening data reveals diverse activity profiles, low global hit frequency, and many consistently inactive compounds. *J. Med. Chem.* **60**, 3879-3886 (2017).

89 Gilberg, E., Stumpfe, D. & Bajorath, J. Activity profiles of analog series containing pan assay interference compounds. *RSC Adv.* **7**, 35638-35647 (2017).

90 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 dataset. *ACS Med. Chem. Lett.* **9**, 792-796 (2018).

91 Senger, M. R., Fraga, C. A., Dantas, R. F. & Silva Jr., F. P. Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discovery Today* **21**, 868-872 (2016).

92 Gilberg, E., Jasial, S., Stumpfe, D., Dimova, D. & Bajorath, J. Highly promiscuous small molecules from biological screening assays include many pan-assay interference compounds but also candidates for polypharmacology. *J. Med. Chem.* **59**, 10285-10290 (2016).

93 Perna, A. M. *et al.* Fragment-based de novo design reveals a small-molecule inhibitor of Helicobacter Pylori HtrA. *Angew. Chem. Int. Ed.* **54**, 10244-10248 (2015).

94 Mike, L. A. *et al.* Activation of heme biosynthesis by a small molecule that is toxic to fermenting Staphylococcus aureus. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8206-8211 (2013).

95 Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today* **20**, 318-331 (2015).

96 Schwaighofer, A. *et al.* Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J. Chem. Inf. Model.* **47**, 407-424 (2007).

97 Rodrigues, T. *et al.* De novo fragment design for drug discovery and chemical biology. *Angew. Chem. Int. Ed.* **54**, 15079-15083 (2015).

98 Cheng, J., Tegge, A. N. & Baldi, P. Machine learning methods for protein structure prediction. *IEEE Rev. Biomed. Eng.* **1**, 41-49 (2008).

99 Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G. & Strother, S. C. Machine learning in medical imaging. *IEEE Signal Process. Mag.* **27**, 25-38 (2010).

100 Zhang, L., Tan, J., Han, D. & Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* **22**, 1680-1685 (2017).

101 Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604-610 (2018).

102 Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **20**, 458-465 (2015).

103 Reker, D., Schneider, P., Schneider, G. & Brown, J. B. Active learning for computational chemogenomics. *Future Med. Chem.* **9**, 381-402 (2017).

104 Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69 (1982).

105 Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273-297 (1995).

106 Weston, J. *et al.* in *Advances in Neural Information Processing Systems 13* (eds T. K. Leen, T. G. Dietterich, & V. Tresp) pp. 668-674 (MIT Press, 2001).

107 Ekins, S. *et al.* Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis. *Mol. Biosyst.* **6**, 2316-2324 (2010).

Figure captions

Figure 1. Mechanisms of unspecific protein inhibition by colloidal aggregates. Poorly water-soluble molecules aggregate at concentration values higher than the critical aggregation concentration. These aggregates can engage in unspecific interactions with proteins, which lead to the protein's inactivation through different mechanisms. Two mechanisms of unspecific protein binding have been reported: (i) formation of large aggregates that sequester and partly denaturate proteins and, (ii) formation of small conglomerates that mimic protein monomers. The example of TNFα inhibition by a conglomerate mimicking a protein scaffold is given (PDB 5MU8)[25]. One should note however that the protein scaffold mimicry mechanism was identified by x-ray crystallography, which not always accurately represents events in solution.

Figure 2. SCAMs are ubiquitous in drug discovery and impact network pharmacology. a) Exemplary structures of natural products and drugs that are commonly used in chemical biology and medicinal chemistry as probes despite aggregating at biochemically relevant concentrations. b) Violin plot of reported bioactivities for quercetin. The activities are biased towards low affinity values, which potentially result in artefactual readouts due to colloidal aggregation. p*Affinity* = –log($K_{i/D}$ or XC_{50}, ChEMBL22). c) Network analysis of potentially erroneous drug target associations extracted from reported bioactivities for experimentally validated SCAMs (p*Affinity* < 4). Each node (vertex) denotes a drug target and associations (edges) are made by a common ligand. The node colour code denotes a connectivity index, *i.e.* from cold (blue; low connectivity) to hot (red; high connectivity) colours. The high network convolution supports multiple drug target relationships that are plagued by artefacts and potentially connects biochemically unrelated drug targets. d) Network analysis of potentially true drug target associations extracted from reported bioactivities for experimentally validated SCAMs (p*Affinity* ≥ 7). The node colour code denotes a connectivity index, *i.e.* from cold (blue; low connectivity) to hot (red; high connectivity) colours. The high network convolution supports multiple drug target relationships that ought to be explored.

Figure 3. Historical evolution of confirmed SCAMs and their underlying scaffolds. The identification of molecules (red) aggregating at biologically relevant concentrations has seen exponential growth in the last years as a result of a few seminal studies focused on the colloidal aggregation phenomenon and the validation of novel assay technology for their detection. The scaffold diversity (blue) in the Shoichet dataset is high, with *ca*. three out five molecules featuring a different

framework. The five most frequent scaffolds, and their counts of occurrences in SCAMs, are highlighted in the blue box. An example of a SCAM for each of the most frequent scaffolds is shown directly above the scaffold in the red box.

Figure 4. Common filters do not detect SCAMs. Pie charts of compound filtering with commonly employed substructural filters. Known aggregating molecules (from the "Aggregator Advisor" database; outer pie charts) and the molecular weight-proportional ChEMBL set of random molecules (inner pie charts) display similar numbers of alerts for widely accepted substructures as computed from a range of filters, implemented by Ekins *et al*.[107] It follows that all substructural filters are insensitive to SCAMs, with the exception of ALARM NMR, which does not appear to be specific. Hit Dexter identifies aggregators but is also not specific. REOS = Rapid elimination of swill[70,72]; GSK = Hann filters[10]; PAINS = Pan-assay interference compounds[7,8,15]; Green = pass filter; Red = fail filter.

Figure 5. Comparison three different *in silico* methods for the automated identification of SCAMs. a) Venn diagram showing overlapping and diverging predictions. The Aggregator Advisor predicts colloidal aggregation through similarity searches with the reference database, *i.e.* it does not involve a training step. Random forests (RF) and support vector machines – recursive feature elimination (SVM-RFE) employ different machine learning algorithms (Box 3) to teach a computer in distinguishing patterns. Despite being constructed with the same reference data including aggregators/non-aggregators, the methods still behave differently when applied prospectively to a random subset of ChEMBL molecules. The consensus prediction of all classification models consists of only 241 molecules, *i.e.* 1.9% of the database size, of which only <1% were known aggregators. Amalgamation of similarity search methods with artificial intelligence may provide an expeditious means to confidently predict SCAMs and nuisance behaviour. b) Multidimensional scaling of predicted liable compounds in the random dataset (Box 1), as identified by the Aggregator Advisor (black), SVM-RFE (cyan) and RF (orange). The projection into two new dimensions (coordinates 1 and 2) is computed from a similarity matrix based on Tanimoto coefficients derived from RDKit's MACCS keys. Data shows that compounds identified by the "Aggregator Advisor" do cluster, which is consistent with a similarity-based approach for flagging chemical matter. Both machine learning methods identify molecules further scattered in the (dis)similarity space, without recognizing privileged clusters, which shows that patterns in molecular properties are learned and identified across a wider range of chemical space.

Tables

Table 1a. Experimental methods for the detection of colloidal aggregates.

	Detergent sensitivity	Dynamic light scattering	Centrifugation	Fluorescence
Assay type	Biochemical	Biophysical	Biophysical	Biophysical
Principle	Enzyme inhibition by SCAMs is disrupted through addition of a detergent (*e.g.* Triton X-100, Tween 80).	Measures the Brownian movement of particles, which correlates with their hydrodynamic radius. It assumes a certain particle shape.	Centrifugation induces the formation of pellets originating from aggregates. Bioactivity change before and after centrifugation is a function of presence of aggregates.	Aggregates change the meniscus curvature in multi well plates, which affects the fluorescence intensity detected by a top read fluorescence plate reader.
Throughput	Depends on assay throughput	Low, but also amenable to multi well plates	Low to high	High in multi well plates
Operational cost	Low per data point in HTS mode	Low per data point	Low	Low per data point in HTS mode
Accessibility	Easy to implement. Gold-standard method using model enzymes.	Easy to implement but requires an expensive equipment. Gold-standard method.	Very easy to implement	Easy to implement
Readout	Enzyme inhibition	Light scattering / hydrodynamic radius	Enzyme inhibition of phenotype change	Fluorescence intensity
Interpretability	Easy	Easy	Easy	Easy

Table 1b. Experimental methods for the detection of colloidal aggregates.

	Surface plasmon resonance	Photonic crystal biosensors	^{1}H NMR	DMSO-perturbed enzymatic assay
Assay type	Biophysical	Biophysical	Biophysical	Biochemical
Principle	A plane-polarized light is reflected by a conducting surface at the interface of two media. Non stoichiometric binding by aggregates affects the angle of diffraction.	Biosensors have a sub-wavelength periodic surface that reflects light. The reflected light is modulated by changes in the refractive index induced by aggregates.	At high concentrations, ^{1}H NMR chemical shifts, multiplicity, shape and intensity change as a result of slow tumbling of aggregates in solution.	High DMSO content generates protein populations – productive and non-productive. Binding of aggregates to the denatured form results in lower inhibitory activity against the fully folded form.
Throughput	Low	High in multi well plates	Very low	High in multi well plates
Operational cost	High cost of consumables	Low cost of consumables	Low per data point	Low per data point in HTS mode
Accessibility	Assay conditions must be finely tuned and requires expensive equipment	Easy to implement, fast, but requires specialized equipment	Easy to implement but requires a very expensive equipment and analysis	Easy to implement
Readout	Refractive index of sample	Refractive index of sample	Chemical shifts of ^{1}H nuclei	Enzyme inhibition
Interpretability	Difficult	Easy	Difficult	Easy

Table 2. Comparison of *in silico* SCAM prediction methods.

	Seidler *et al.*[30]	Feng *et al.*[32]	Rao *et al.*[81]	Aggregator Advisor[51]
Algorithm	Recursive Partitioning	Random Forest	Support-Vector Machine	Chemical similarity
Training / model complexity	Low	Medium	High	None
Interpretability	High	Medium	Low	By chemistry experts
Data quantity	Low	High	High	Very high
Advantage	Good Interpretability	High precision	High recall	Continuously add data
Disadvantage	Low accuracy	Low recall	Low precision	No generalization
Applicability domain	Narrow[a]	Wide	Wide	Narrow
Prediction speed	Very fast	Fast	Fast	Medium

[a] as originally implemented with a small training set.

Figure 1

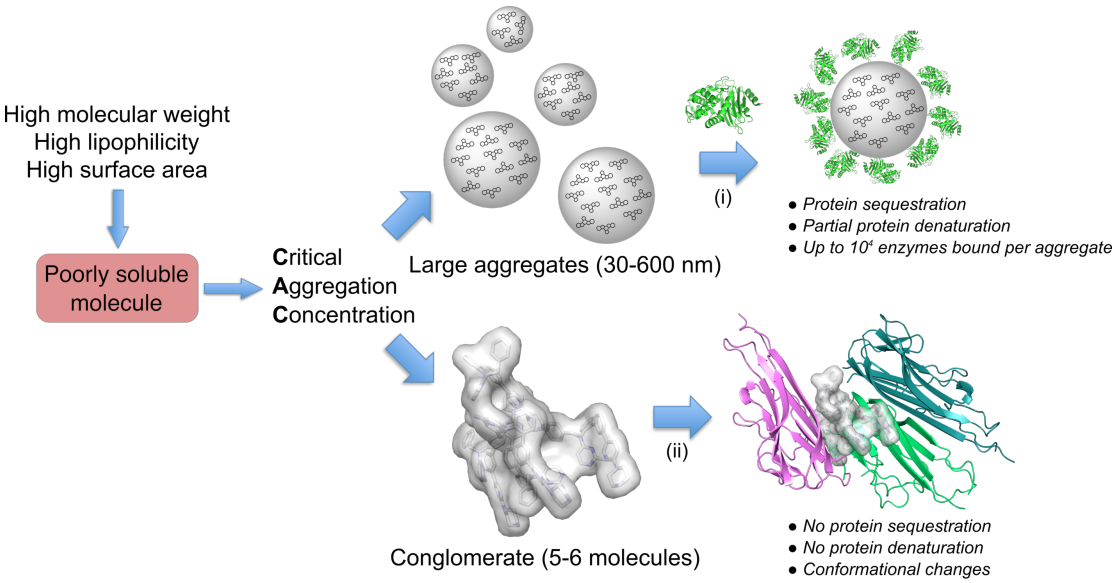

High molecular weight
High lipophilicity
High surface area
Poorly soluble molecule
Critical
Aggregation
Concentration
Large aggregates (30-600 nm)
(i)
• Protein sequestration
• Partial protein denaturation
• Up to 10⁴ enzymes bound per aggregate
(ii)
Conglomerate (5-6 molecules)
• No protein sequestration
• No protein denaturation
• Conformational changes

Figure 2

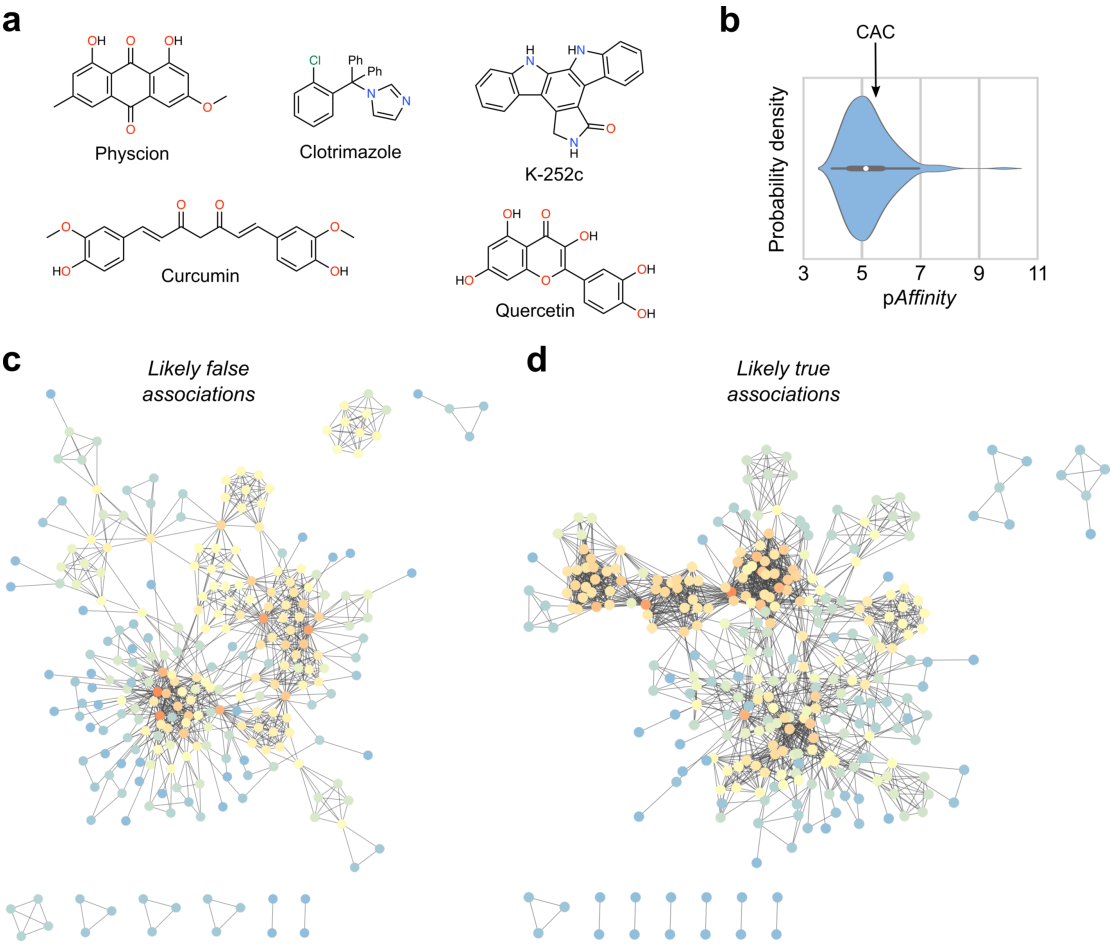

a
Physcion
Clotrimazole
K-252c
Curcumin
Quercetin
b
CAC
Probability density
3
5
7
9
11
pAffinity
c
Likely false associations
d
Likely true associations

Figure 3

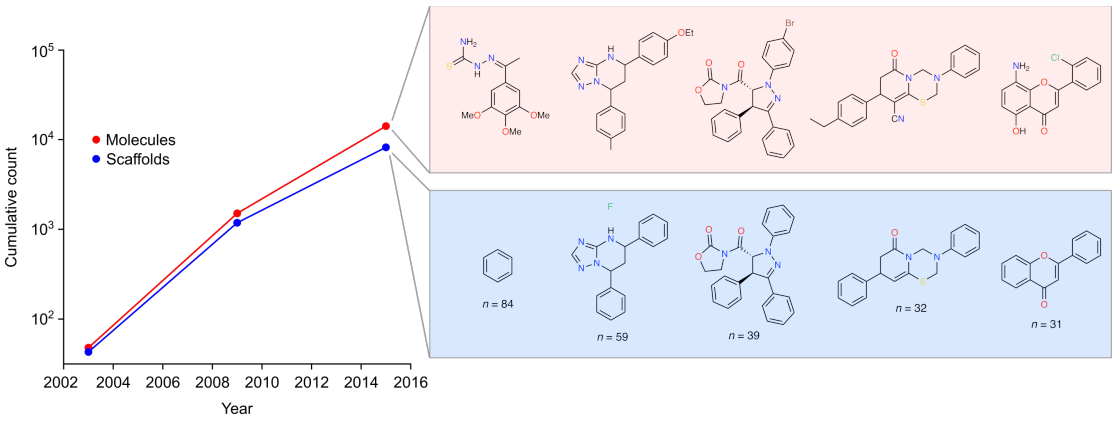

Cumulative count
10^5
10^4
10^3
10^2
Molecules
Scaffolds
2002
2004
2006
2008
2010
2012
2014
2016
Year
n = 84
n = 59
n = 39
n = 32
n = 31

Figure 4

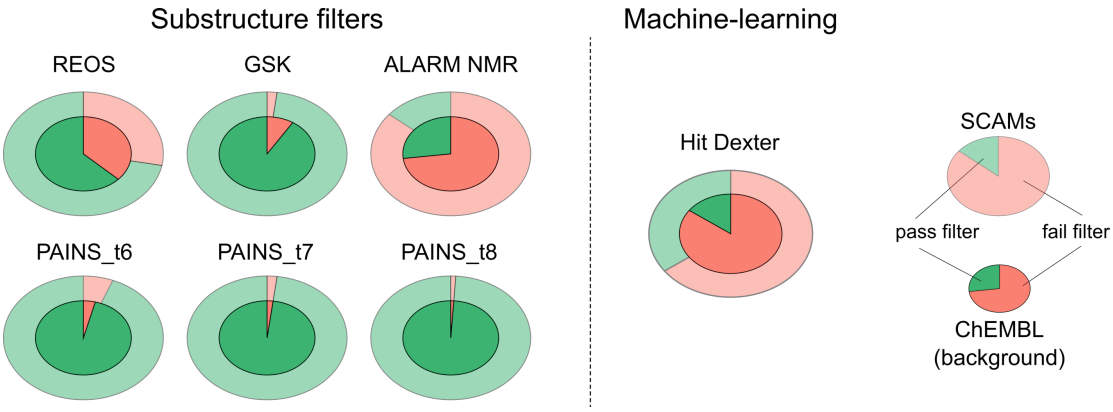
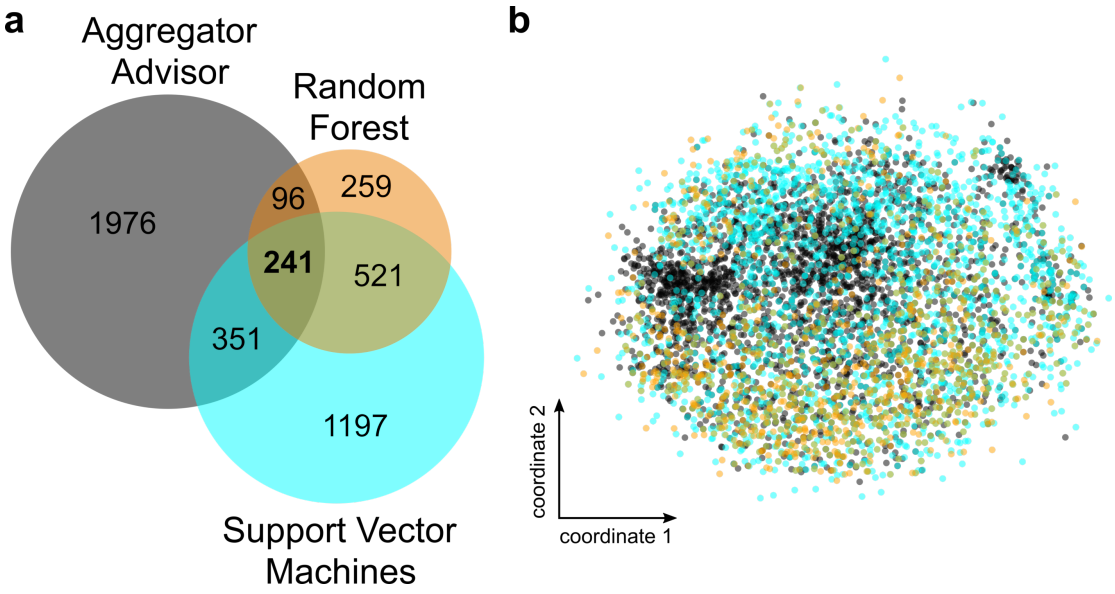
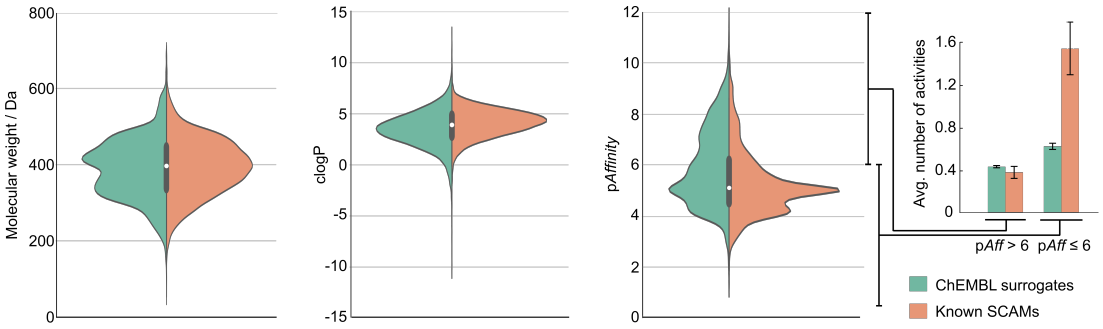

Substructure filters
Machine-learning
REOS
GSK
ALARM NMR
PAINS_t6
PAINS_t7
PAINS_t8
Hit Dexter
SCAMs
pass filter
fail filter
ChEMBL
(background)

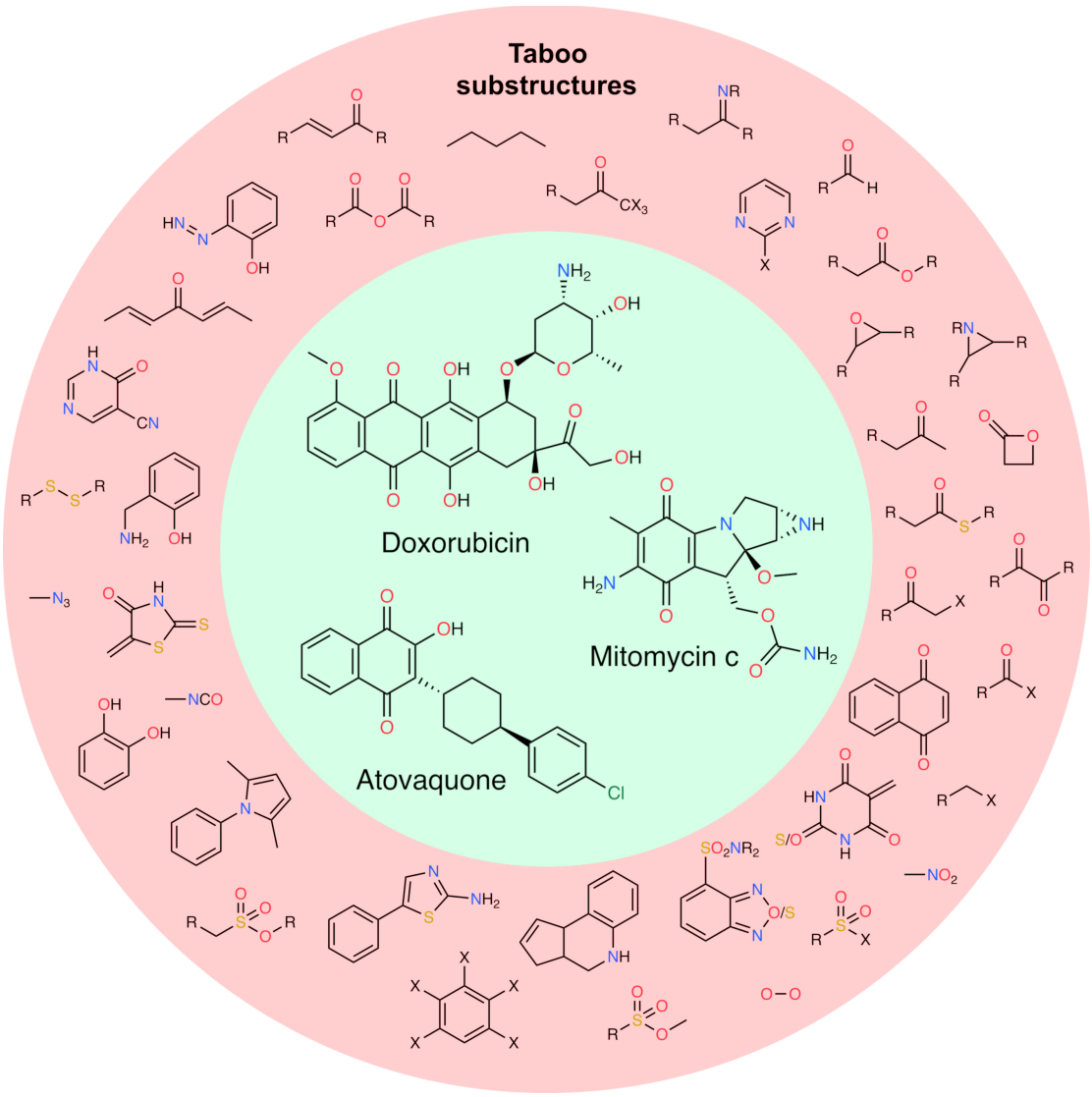
Figure 5

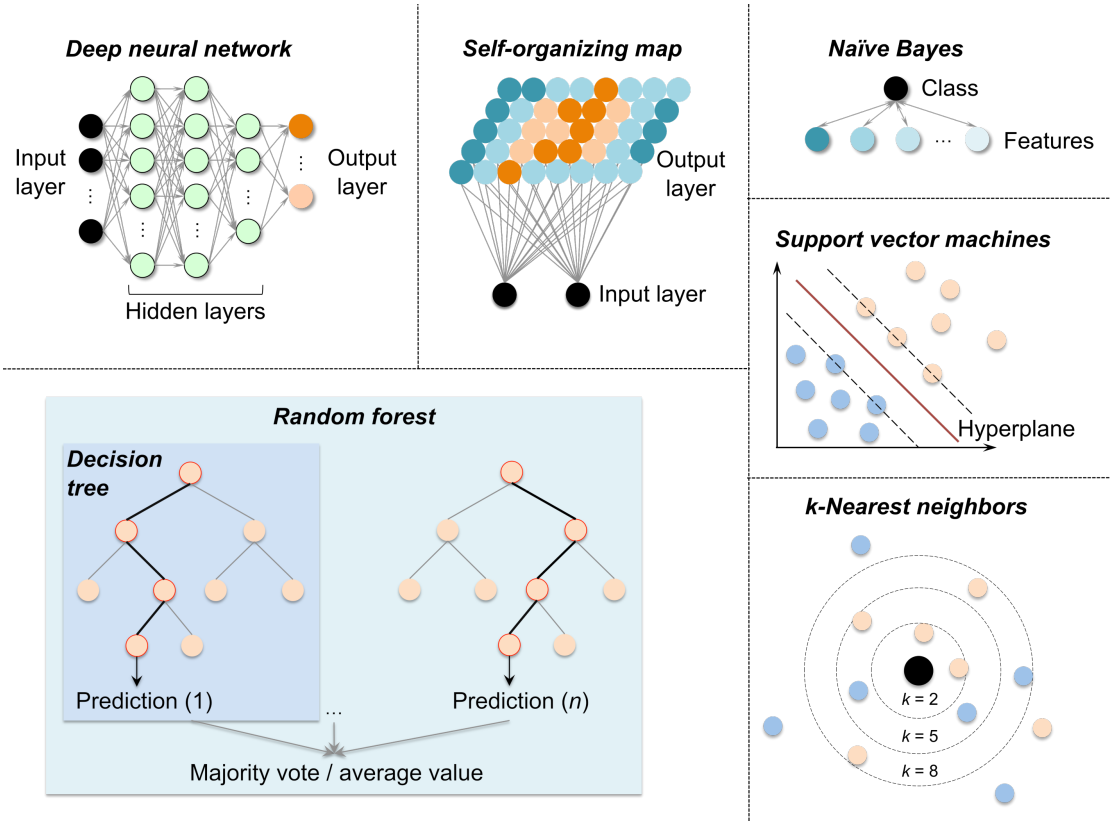

a
Aggregator Advisor
Random Forest
Support Vector Machines
1976
96
259
241
521
351
1197
b
coordinate 2
coordinate 1

Box 1


Molecular weight / Da
800
600
400
200
0
clogP
15
10
5
0
-5
-10
-15
pAffinity
12
10
8
6
4
2
0
Avg. number of activities
1.6
1.2
0.8
0.4
0
pAff > 6
pAff ≤ 6
ChEMBL surrogates
Known SCAMs

Box 2




Deep neural network
Input layer
Output layer
Hidden layers
Self-organizing map
Output layer
Input layer
Naïve Bayes
Class
Features
Support vector machines
Hyperplane
Random forest
Decision tree
Prediction (1)
Prediction (n)
Majority vote / average value
k-Nearest neighbors
k = 2
k = 5
k = 8