Combating small-molecule aggregation with machine learning

Poor solubility is the prime source of false-positive hits in drug discovery. Lee et al. implement and evaluate DeepSCAMs to identify molecules aggregating at a typical screening concentration. They estimate that up to 15%–20% of all small molecules are prone to aggregation at 30 μM.
Article

Combating small-molecule aggregation with machine learning

Kuan Lee,1,6 Ann Yang,1,6 Yen-Chu Lin,1,* Daniel Reker,2 Gonçalo J.L. Bernardes,3,4 and Tiago Rodrigues1,5,7,*

SUMMARY

Biological screens are plagued by false-positive hits resulting from aggregation. Methods to triage small colloidally aggregating molecules (SCAMs) are in high demand. Herein, we disclose a neural network to flag such entities. Our data demonstrate the utility of machine learning for predicting SCAMs, achieving 80% of correct predictions in an out-of-sample evaluation. The tool is competitive with a panel of expert chemists, who correctly predict 61% ± 7% of the same molecules in a Turing-like test. Our computational routine provides insight into features governing aggregation that had remained hidden to expert intuition. Further, we quantify that up to 15%–20% of ligands in publicly available chemogenomic databases have high potential to aggregate at a typical screening concentration (30 μM), imposing caution in systems biology and drug design programs. Our approach provides a means to augment human intuition and mitigate attrition and a pathway to accelerate future molecular medicine.

INTRODUCTION

Drug discovery is fueled by small molecules, either as tools to interrogate biology or as leads for investigational therapeutics.1–3 The successful development of such entities entails a transformative power on disease modulation, but a large proportion of them fail to reach the clinic because of attrition.4 In that regard, it is increasingly recognized that a significant fraction of screening molecules present low aqueous solubility and form nano- or microscale agglomerates. These colloidal aggregates can bind unspecifically to proteins, inducing local denaturation and apparent inhibition.5 Although undirected interactions between small molecules and proteins are not desirable, most biological assays lack resolution to distinguish between “pathological” and “directed” target engagement, thereby incorrectly presenting the aggregating molecule as a positive result. Indeed, colloidal aggregates account for the largest source of false-positive hits in high-throughput screens, surpassing other well-documented “con artists,” such as the pan-assay interference compounds.6,7

Vendor molecules, approved drugs, and natural products are used to query biology on a systems level. Given the widespread potential for small-molecule promiscuity8–10 and aggregation,8,11–13 assay recommendations have been made. Those recommendations allow scrutinizing screening hits and de-prioritize chemical matter, displaying unspecific target modulation.14 In particular, biochemical assays, with and without detergent, and dynamic light scattering (DLS) have become the gold-standard technologies for interrogating aggregation14,15 and detect a fraction of

---

1Insilico Medicine Taiwan Ltd., Taipei City 110208, Taiwan
2Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA
3Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
4Instituto de Medicina Molecular, Faculdade de Farmácia, Universidade de Lisboa, Avenida Professor Egas Moniz, 1649-028 Lisbon, Portugal
5Instituto de Investigação do Medicamento (iMed), Faculdade de Farmácia, Universidade de Lisboa, Avenida Professor Gama Pinto, 1649-003 Lisbon, Portugal
6These authors contributed equally
7Lead contact
*Correspondence:
jimmy.lin@insilicomedicine.com (Y.-C.L.),
tiago.rodrigues@ff.ulisboa.pt (T.R.)
https://doi.org/10.1016/j.xcrp.2021.100573
molecules with promiscuous behavior. Whereas biochemical assays can be expensive and laborious, DLS is more accessible and allows determining the dynamic radius of particles in solution at a given concentration. Still, it rapidly becomes cumbersome, from both execution and data analysis perspectives, when scaled. As an alternative, the pharmaceutical community is seeking heuristics to expedite the detection of liable chemical matter at the primary screen stage.

“Drug-likeness” and the “quantitative estimate of drug-likeness” (QED) have been widely used to prioritize small molecules in early discovery programs. Although valuable, these composite metrics were developed for disparate purposes and do not correlate with attrition. Many approved drugs fail to comply with at least one of Lipinski’s drug-like rules for oral absorption, and the QED overlooks the background distribution of molecular properties for non-drugs. Therefore, methods that are tailored to flag promiscuous small molecules are in high demand. If accurate, those tools may provide practical research companions, limit laborious and expensive experimentation, streamline the (de)-validation of screening hits, and accelerate development pipelines by diminishing attrition at all stages. In the past two decades, several machine-learning methods have been devised with the goal of expediting the identification of frequent hitters in a data-driven fashion. From neural networks to extra trees (HitDexter), random forests, Bayesian inference (BadApple), \(k\)-nearest neighbors (kNN), and support vector machines (SVMs), a multitude of algorithms have been successfully implemented to triage screening molecules. Furthermore, associative neural networks, as in the Luciferase Advisor tool, have shown high balanced accuracy in retrospective studies toward the identification of interference compounds in a specific reporter gene bioassay. Although valuable, all these algorithms do not hint on a molecular basis for interference, which can be due to auto-fluorescence, reactivity, aggregation, and/or other factors.

The Aggregator Advisor provides a solution for a better understanding of the underlying interference mechanism by flagging small colloidally aggregating molecules (SCAMs) on the basis of chemical similarity and lipophilicity computations. Still, one may argue that this tool presents limited applicability on a wider scale because of an intrinsic inability to extrapolate chemical patterns. Similarity searches in the Aggregator Advisor provide insufficient generalizability, despite the available reference data (currently >12,500 structures). In practice, many SCAMs will pass the filter unnoticed. Nonetheless, its source data have been used to build more sophisticated flagging systems, such as an SVM model, the SCAM Detective, and ChemAGG, which provide new vistas on colloidal aggregation. Although valuable, one limitation of these methods is that they do not account for molecule concentration in the assay, which we reason as of paramount importance for an appropriate prediction contextualization. To that end, we here develop and evaluate a generally applicable neural network that more realistically gauges the impact of colloidal aggregation in small-molecule discovery programs. We show that by focusing on low but homogeneous data—with regard to screening medium identity, molecular concentration, and colloidal aggregation annotations—we can boost predictive accuracy to 80%. The machine-learning tool is competitive with available computational technologies that predict a similar endpoint and outperformed a small panel of expert chemists in an out-of-sample evaluation. Moreover, using state-of-the-art model interpretation, we identified both hitherto unknown chemical features correlated with molecular aggregation and pitfalls in the formalization of domain knowledge. Finally, a large literature screen revealed >1.1 million potentially erroneous ligand associations that may skew pharmacology networks at a small-molecule
concentration of 30 μM. Our predictive pipeline may contribute to advance future molecular design by alleviating attrition and thus reliably assist expert intuition in decision making.

RESULTS AND DISCUSSION

Machine learning for flagging SCAMs

We built a fully connected deep learning classifier from readily available DLS data. We elected to train a model based on this dataset because of its homogeneity and causal link between aggregation and a specified and typical primary screen concentration (30 μM). The relationship between test concentration and label is not considered in previous tools. Although the Aggregator Advisor dataset is more comprehensive, it would provide an unsuitable training set in the present case. This is due to encompassing data from different sources (e.g., DLS and detergent-based biochemical assays) and, more important, molecular annotations at different concentrations, which do not map onto our measured endpoint. Together, these are expected to affect model performance and afford predictions suited to a distinct research question (i.e., wherein screening concentration is not relevant). Aggregation and pathological target modulation might result from assay-specific components and can be identified by running paired biochemical assays (with and without detergent). However, prior contributions have shown an excellent correlation between unspecific target binding and particle formation under the DLS screening conditions used in our study. We thus deemed the DLS readings appropriate for modeling and a motivated proxy to estimate promiscuous binding.

Curation of the DLS data (i.e., elimination of molecules originally labeled as “ambiguous”) provided the required training set, comprising 916 entities annotated with an “aggregation”/“non-aggregation” readout (30:70 ratio, respectively) at a typical primary screen concentration (30 μM). We calculated substructural Morgan fingerprints (2,048 bits, radius 3) and topological physicochemical properties (n = 199; RDKit) for each molecule to abstract chemical connectivity and obtain a high-dimensional descriptor. We then implemented an exhaustive parameter search routine to identify a classifier with optimal performance (cf. Experimental procedures; Figures S4 and S5). As assessed through stratified 10-fold cross-validation studies, a feed-forward neural network with three hidden layers (DeepSCAMs; Figures 1A and 1B) that could correctly retrieve molecules from both classes (Figure S1) was selected for downstream studies. In parallel, we ran adversarial controls to de-validate the soundness of our classifier. By randomizing the target class annotations and subsequently building an optimized neural network, we observed that model performance decreased and became comparable with random guessing (Figure 1B). This supports the disruption of relevant patterns during the Y-shuffling process, the meaningfulness of the abstractions used, and the potential utility of our native model. In part, these results rule out overfitting, which is a realistic concern for neural networks harnessing high-dimensional descriptors with small yet real-world datasets.

To prospectively challenge the utility of our machine-learning model, we queried >1.2 million purchasable screening molecules and predicted their probabilities of generating colloidal aggregates. For experimental evaluation, candidate selection was performed by using the MaxMin algorithm to ensure a chemotype-agnostic, diverse selection while covering the whole range of prediction probabilities as either “aggregators” or “non-aggregators” (cf. Table S1). We procured 65 screening molecules and executed DLS screens to monitor both particle formation and
auto-correlation curves at a concentration of 30 μM (Figure 1C). On the basis of the collected data, we confirmed 80% of correct predictions, which is fully aligned with the cross-validation studies. Not only did DeepSCAMs present near perfect performance for high-confidence predictions (19 of 20 correct predictions when class probability was >95%) but also good performance for lower confidence outputs (33 of 45 correct predictions), supporting the ability to extrapolate learned patterns. More specifically, 77% (24 of 31) and 82% (28 of 34) of SCAMs and non-SCAMs, respectively, were correctly identified in this prospective evaluation. Among the 13 mispredictions, 10 (77%) presented label assignment probabilities similar to 50%, which fully rationalizes the computational outcome. Importantly, the average
Tanimoto index between the 65 screening molecules and their nearest neighbors in the DeepSCAMs training set is only 0.20 ± 0.09 (Figure 1D). An arbitrary Tanimoto index cut-off can be set between 0.65 (e.g., Szilágyi et al.42) and 0.8 (e.g., Irwin et al.36). In all our prospective evaluation molecules, the similarity to the nearest neighbor falls under the cut-off. Furthermore, no apparent relationship between the similarity index and predicted class probability was found (cf. Figure S3). We also noted that the selected chemotypes extend beyond the typical “drug-like” space in the training set (QED_evaluation_molecules = 0.52 ± 0.18 [n = 65] versus QED_training_set = 0.63 ± 0.18 [n = 916]; p value = 2.28 × 10⁻⁵; Welch t test).22 Together, these observations attest to the substructural dissimilarity between the evaluation and training sets, motivate the former as non-trivial use cases from a machine-learning vantage point and thus are appropriate for challenging the generalizability and domain of applicability of DeepSCAMs.

We next contextualized the utility of DeepSCAMs by using six optimized machine-learning algorithms on our training data (kNN, decision trees, random forests, AdaBoost, naive Bayes, and SVM; cf. Tables S2–S4). Although we found that those alternative methods did not present predictive performance justifying their use over DeepSCAMs (cf. Tables S3 and S4; Figure S2), the results equally suggest that other reasonable approaches may be developed. Of note, our method was superior relative to the baseline flagging tool, Aggregator Advisor (cf. Table S1). While DeepSCAMs correctly classified a large proportion of the confirmed aggregators in the out-of-sample evaluation set, the Aggregator Advisor retrieved many false negatives (SCAMs correctly identified: 76% for DeepSCAMs versus 29% for Aggregator Advisor; Figure 1E; p value= 0.05 for Aggregator Advisor versus p value= 1.17 × 10⁻⁶ for DeepSCAMs, two-sided binomial test against random label selection). The result highlights an important caveat of similarity searches; they depend on the number and diversity of reference molecules, thus leading to a large fraction of false negatives if the chemical space is not broadly covered. Finally, our model also compared favorably with ChemAGG (Figure 1F; Table S1; p value = 2.62 × 10⁻³, two-sided binomial test),39 the SCAM Detective, and an SVM for aggregate prediction (Table S1). We hypothesized that the enhanced generalizability of DeepSCAMs relative to other reported methods is linked to a leaner (<10% in size to ChemAGG, SCAM Detective, and SVM method) and more homogeneous training set. Because the Aggregator Advisor reference set is heterogeneous with regard to the aggregating concentration and screening technologies, we deemed it inappropriate for external evaluation of our method and use as a benchmark dataset. To test this, we instead trained DeepSCAMs with the Aggregator Advisor data to investigate if diverging aggregating concentrations could affect the resulting model performance. Given that the Aggregator Advisor data are highly imbalanced (95% aggregators), we down-sampled the majority class to obtain an even distribution of examples in a total of 653 molecules. Training led to a classifier with accuracy of 75% (n = 3), which could be decreased by shuffling the target variables (50%, n = 3). These control computations show that curating data with an experimental physical-organic chemistry motivation is desirable and an advantageous strategy that had previously been overlooked. As such, the tailored training set of DeepSCAMs provides a viable basis for augmenting machine perception25 and a preferential means of answering a research question that cannot be tackled otherwise.

**Chemical knowledge through model interpretation**

Despite the prospective utility and favorable performance of DeepSCAMs relative to other in silico methods, we wondered if our software tool could correctly formalize domain knowledge and augment perception on established, yet unwritten rules of
chemical intuition. We thus asked 15 PhD-level medicinal chemists and chemical biologists from academia and industry in Europe, Asia, and the United States to benchmark the tool, as previously conducted by us40 and others.44–46 In a Turing-like test, the 15 experts were instructed to label each of the 65 evaluation set molecules as “SCAM” or “non-SCAM,” considering a test concentration of 30 µM in buffered aqueous solution (pH 7). The results show that DeepSCAMs has a competitive vantage point relative to the chemists and efficiently sieved through hidden chemistry patterns. Specifically, DeepSCAMs was not only able to correctly classify cases that were consensual for the expert panel but also others in which the majority of the chemists failed (Figures 2A and 2B). Interestingly, the observed general disagreement between the surveyed experts with distinct clusters of domain knowledge reiterates how subjective and intuition based the prediction of aggregation is (cf. Figure S10). We do note that the short survey performed herein is merely indicative and would require a larger panel of chemists and molecules to allow broader generalizations supported by statistics. However, the data suggest that DeepSCAMs offers a robust, unbiased and efficient solution to flag (non)-SCAMs beyond intuitive rules.

To gauge the impact of each molecular feature on the DeepSCAMs output and shed light on the prediction disparities relative to expert intuition, we calculated Shapley values from game theory, as a solution maintaining both local accuracy and consistency. In short, Shapley values quantify the cooperative impact of a given feature as the change in the expected value to the model’s output when a feature is observed versus the unknown47 (Note S1). A similar approach has been successfully used to generate knowledge in different chemical problems.48–50 Deconvolution of our machine-learning model into such quantitative metrics showed the multidimensional nature of small-molecule aggregation and motivated (dis)agreements with the expert chemists’ predictions (Figures 2B and 2C). Specifically, it considered that sulfonamides can promote solubility and that flexible and lipophilic amines can provide the blueprints for aggregation, a realization that was either overlooked or unnoticed, as only one of the chemists was able to correctly classify the molecule in question (Figure 2B). We challenged the relevance of this hitherto unknown trend through repeated SHAP analyses and the use of LIME51 as a complementary approach (cf. Figures S6–S9) to provide statistical robustness in the model interpretation. While ranking features by importance is a viable means to interpret the decision function, it is then desirable to evaluate them and search for a correlation with physical phenomena. With that in mind, we purchased and screened three analogs of a SCAM that contained the piperidylmethanamine scaffold. This motif had been critical for prediction of the reference molecule as a SCAM (cf. Figure 2B). Through DLS, we confirmed non-aggregation in all three cases, which counters intuition, considering the molecular similarity to our reference SCAM. The result is interesting because DeepSCAMs was able to confidently predict the DLS outcome, despite the presence of the piperidylmethanamine substructure (cf. Figure S11). Ultimately, it supports that establishing a causal link between structural features and aggregation is challenging and requires a large-scale study on its own to comprehensively evaluate the important features. In a couple of small molecules, both our tool and the majority of consulted experts failed to recognize patterns leading to either aggregation or non-aggregation (Figure 2B). The result shows there are a number of relevant aggregation signatures that remain unknown; these can be unraveled further, as data in contiguous search spaces become available. The same reasoning can be applied to the nonsensical predictions we obtained for more established test cases, in which a significant proportion of experts were able to correctly assign a label. For some of those molecules, the DeepSCAMs prediction did not deviate significantly from random guessing (i.e., near misses) but still failed to recognize molecular shape
and conformations that could promote one or the other class. Despite the marginal failures discussed herein, the high overall accuracy in out-of-sample predictions attests to the value of DeepSCAMs, wherein the model can establish data correlations that are currently unexpected or unapparent.

From a model-wide vantage point, we identified that the number of relevant features largely exceeds the typical number, four, efficiently processed by humans. This may help explain the competitive performance of DeepSCAMs. Further, our data corroborate that substructural fingerprints are appropriate descriptors to identify SCAMs. Some bits are highly positively or negatively associated with the probabilistic output (Figures 2B and 2C), despite the potential interpretability caveats and dependence

Figure 2. Formalization of expert intuition

(A) Prediction comparison between DeepSCAMs and 15 expert chemists from academia and industry. The top row denotes the experimental result (ground truth) from the DLS screen at 30 μM (pH 7), followed by DeepSCAMs and the researchers’ answers; gray, SCAM; white, non-SCAM; red, incorrect prediction; green, correct prediction; asterisk, missing answer. The percentage of correct predictions is provided as well as a statistical analysis versus random guessing (two-sided binomial test).

(B) Structure of selected molecules that were correctly or incorrectly predicted as SCAM or non-SCAM by both DeepSCAMs and expert chemists. In multiple cases, the algorithm and chemists are in agreement and in other cases in disagreement. For predictions correctly made by both software and chemists, the relevant structural features are highlighted in red (supporting SCAM) and green (supporting non-SCAM). For a correct prediction by DeepSCAMs in disagreement with the expert panel, the relevant structural features are highlighted in orange. Collectively, those features are underrepresented in the training set (two molecules), which may explain the responses given by the chemists. For molecules mispredicted by DeepSCAMs but correctly labeled by the experts, the highlighted motifs represent consensual features as per human intuition (n = 5).

(C) Model interpretation (on the basis of training data) showing the top five global features using Shapley values. The top two physicochemical properties are also provided as well as their global ranks in parenthesis. Data show that different fingerprint bits have varying weights on the model output (prediction as SCAM) and that both a high number of aliphatic carbocycles and molecular partial charges are negatively correlated with prediction as SCAM.

(D) Isomap of activation outputs for each of the three layers in DeepSCAMs. The projection shows that training data become better separated with the neural net depth. Prospective examples that were incorrectly labeled by DeepSCAMs but not by the majority of surveyed chemists are highlighted (black dots). For details on how to distinguish well- from poorly separated data, see Figures S4 and S5 and Tables S2–S6.
on the algorithms used. Curiously, calculated logP, a property of paramount importance for the Aggregator Advisor, did not rank among the top 20 features in this instance. However, the result does not de-validate the importance of logP, which is expected to play a more or less important role for explaining individual predictions. Rather, it demonstrates that flagging molecules on the basis of a single calculated physicochemical descriptor may be an over-simplistic approach, leading to mispredictions. Conversely, molecular partial charges and the number of aliphatic carbocycles more significantly affect model performance. For example, a large number of aliphatic carbocycles lowers the probability of a molecule being predicted as a SCAM (Figure 2C). This aligns with chemical intuition, as decreased stacking interactions promote aqueous solubility. For enhanced interpretability, we next projected the decision path of DeepSCAMs via the corresponding activation outputs per layer. The projections reveal that SCAMs and non-SCAMs in training (916 molecules) and external evaluation (65) sets become increasingly well separated along the neural network depth, despite originally overlapping in the input space (Figure 2D; Figures S5 and S6). This not only depicts the complexity of the problem in hand, wherein both classes are naturally intermixed, but also justifies the chosen model architecture and rationalizes the prospective utility of the tool. A close inspection of the projected data revealed opportunities for method development. For example, some molecules were poorly separated from the opposite-class neighbors, irrespective of being correctly or misclassified (cf. Figures S4 and S5; Tables S3 and S4). Overall, our approach provides an intuitive and readily visualizable means to rationalize the ability of DeepSCAMs in formalizing chemical knowledge (cf. selected molecules in Figure 2). Because we were able to delve into the inner workings of DeepSCAMs, build trust in its architecture, and obtain interpretable and explainable decision processes that are largely aligned with expert intuition, we next expanded the application of our tool to bioactive compound databases in the hope of illuminating liable matter.

Identification of biases in the literature

Previously, about 7% of all compounds in the medicinal chemistry literature had been flagged by the Aggregator Advisor. Although experimentally unconfirmed, such a large fraction of nuisance compounds can perturb exploratory and development pipelines. Flagging molecules as early as possible is desirable and can alleviate the pursuit of unfruitful research lines. However, the lack of concentration-effect interdependence in the Aggregator Advisor output, as discussed above, makes gauging the impact of aggregation in primary screens currently inaccessible. Considering the higher generalizability of DeepSCAMs and its correlation with a primary screen concentration, we re-interrogated the prevalence of potential aggregators in >1.8 million ChEMBL molecules. The goal here was to provide machine-learned insights onto SCAMs across the chemical biology literature, when and if used in assays at 30 μM, which is a realistic scenario in primary screens aiming at identifying hits for medicinal chemistry elaboration. While accounting for the predictive uncertainty in DeepSCAMs, we found that up to 15%–20% of all chemical entities are anticipated to aggregate at 30 μM (Figure 3) and can mislead biological investigations. The result is 2- to 3-fold higher than the prior estimate but is aligned with data from a small random screen. In a more extreme case, 95% of screening hits have shown pathological target engagement, including not only colloidal aggregation but also auto-fluorescence and reactivity with assay components. Our computations thus confidently support a higher prevalence of SCAMs than previously estimated. They also provide a data-motivated interpretation for the qualitatively known, but often neglected bias toward nuisance compounds in molecular medicine. Apparently, this is an established and deep-rooted trend, as computed by DeepSCAMs in a time-series analysis of ChEMBL 2009–2019 (Figure 3).
Structurally, the predicted SCAMs have a high degree of polyaromatic moieties (97% of all molecules), which endorse an impaired kinetic solubility. One possible explanation for this observation is the ease of synthesis of such motifs relative to \( sp^3 \) hybridization-rich frameworks through C-C bond formation reactions, such as Suzuki-Miyaura coupling-type chemistry. The trend was corroborated by analysis of the BindingDB, wherein 20% of the database could be predicted to aggregate at 30 \( \mu \)M while using our learning algorithm.

Given the use of databases such as ChEMBL and BindingDB in model implementation and analyses of structure-activity relationships, this result is alarming for the medicinal chemistry and chemical biology communities, as SCAMs might misguide model development or hit and lead discovery programs. We next extracted protein and activity annotations for each entity, as reported in ChEMBL, to study the impact of predicted SCAMs at an integrated biology level. Our analyses suggest that all protein families are likely affected by false-positive hits and that 5,733 predicted SCAMs have bioactivity annotations—against 531 distinct targets—above 30 \( \mu \)M. Our analyses also showed multiple high IC\(_{50}/EC_{50} \) or \( K_D/K_i \) values for molecules screened against isoforms of cytochrome P450 (Figure 4A). Strikingly, the microtubule-associated protein tau and beta-secretase 1 are common targets of potential SCAMs. Both macromolecules have been explored in Alzheimer’s disease, and the inclusion of entities with a questionable medicinal chemistry signature as prototypes for development might partly explain a number of unsuccessful early discovery programs in this space, together with convoluted biology. From a broader perspective, a total of 6,078 ligand-target relationships are likely compromised, which may incorrectly link 659 distinct targets (Figure 4B). Furthermore, >1.1 million predicted SCAM relationships endorsed by drug target commonalities were identified and, thereby, can inflate their respective ligandable spaces. Undoubtedly, “false hits” and assay interference are established concepts, but quantifying the impact of aggregation on discovery chemistry has been elusive, to the best of our knowledge. Despite the accuracy of DeepSCAMs, the data insights would require large-scale experimental confirmation. Still, our machine-learning tool frames the colloidal aggregation problematic in a way that has profound implications per se. Overall, it reinforces an urgent need for enforcing standardized data quality measures. The implemented measures must be widely adopted, including routine biochemical counter-screens in the presence and/or absence of detergents or DLS experiments. Notoriously, similar recommendations have been proposed, but relatively narrow adoption seems insufficient to revert the observed trends. Our quantitative data resonate with those calls and motivate the development of higher quality probes and drug leads, as the means to gradually reduce the impact of SCAMs in discovery sciences and more efficiently advance exploratory and late-stage development pipelines.
In summary, we built and prospectively evaluated a deep neural network to flag SCAMs. While concentration and screening medium identity are key in governing the colloidal aggregation of small molecules, all available in silico technologies neglect, to the best of our knowledge, their influence toward the binary output. Conversely, DeepSCAMs leverages homogeneous data and indirectly formalizes said physical-chemical correlation to reveal the potential utility of machine learning in this space. In a challenging prospective evaluation, our tool was competitive with preceding computational workflows and a small panel of expert chemists. Yet DeepSCAMs might be improved further as more standardized DLS data are collected. Importantly, we harnessed machine learning to generate chemical knowledge. The statistical model enlightened chemical features, such as lipophilic amines, that might more broadly influence small-molecule aggregation and have so far remained overlooked by most expert medicinal chemists. Moreover, DeepSCAMs exposed persistent and alarming biases in the literature that favor liable chemical matter at high screening concentrations. If experimentally uncontrolled for, those SCAMs can misguide rational drug design across all target families, leading to higher than envisaged consequences in present and future small-molecule development. With the advent of automated, data-driven research workflows in drug discovery, we expect the propagation of misleading structure-activity correlations to disrupt development pipelines and slow down the identification of life-changing therapies. We caution that our data and conclusions refer to primary screens ran at a concentration of 30 \( \mu \text{M} \). A lower percentage of aggregators is expected to be found at other typical screening concentrations (e.g., 10 \( \mu \text{M} \)), which we are currently not able to estimate because of a lack of reference data. We envisage that DeepSCAMs and related tools
can play a crucial role in augmenting domain intuition, probabilistically informing decision-making and safeguarding the design of bioactive matter, thereby supporting next-generation molecular medicine.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact
Further information and requests should be directed to the lead contact, Dr. Tiago Rodrigues (tiago.rodrigues@ff.ulisboa.pt).

Materials availability
All small molecules were obtained through vendor libraries. No new materials were generated in this research.

Data and code availability
All of the data supporting the findings are presented within the article and the Supplemental information. Details on DeepSCAMs workflow are provided in the main text and Supplemental experimental procedures. The code can be accessed at https://github.com/tcorodrigues/DeepSCAMs.

Machine learning and data analyses
All software was implemented in Python 3.7, using the NumPy 1.11.3, Pandas 0.19.2, RDKit 2020.03.1, and Scikit-learn 0.18.1 libraries. DeepSCAMs was constructed with publicly available DLS data,\textsuperscript{15} using a full grid hyperparameter search (alpha = 0.0001/0.001; iterations = 200/500/1,000; activation = logistic/tanh; solver = lbfgs/sgd; learning rate = constant/adaptive; hidden layers = 1–3 layers with 10/100/1,000 neurons) and stratified 10-fold cross-validation for model selection. The selected feedforward neural network uses three hidden layers with 100, 1,000 and 1,000 neurons, respectively, an alpha value of 0.0001, tanh as an activation function, stochastic gradient descent, and a constant learning rate over 200 iterations. All remaining hyperparameters were used as defaults. Data analysis pipelines were performed with SciPy 1.2.3 and plotted in Matplotlib 1.5.3, Matplotlib-Venn 0.11.5, or Seaborn 0.10.1. Figures were compiled in Inkscape 0.91. Computations were carried out on an Apple Mac Pro machine (3.5-GHz processor, 32 GB RAM) or Google Colab. The Chembridge CORE/ExpressPick, ChEMBL v1–25,\textsuperscript{55} and BindingDB\textsuperscript{58} libraries were queried with DeepSCAMs. Note that although “aggregation” measured through DLS can be used to flag promiscuous behavior (and has shown good correlation with biochemical assay outputs\textsuperscript{36}), it does not ensure pathological target engagement per se. Specifically, DeepSCAMs focuses on predicting DLS outputs (i.e., aggregation in a physiologically relevant aqueous medium) and does not account for aggregation induced by assay conditions, such as the presence of specific biomacromolecules. See also Supplemental experimental procedures.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2021.100573.

ACKNOWLEDGMENTS

G.J.L.B. is a Royal Society University Research Fellow (URF) (URF/R1/180019). T.R. is an Investigador Auxiliar, supported by FCT Portugal (CEECIND/00684/2018). We thank V. Lunot for valuable discussions.
AUTHOR CONTRIBUTIONS
K.L., A.Y., and D.R. contributed with machine-learning models and data analyses. Y.-C.L. and G.J.L.B. analyzed data and supervised research. T.R. implemented DeepSCAMs, performed screens, and analyzed data. T.R. conceived, designed, and coordinated the study. T.R. wrote the manuscript, with contributions from the remaining authors. K.L. and A.Y. contributed equally to this study. All authors agreed on submitting the current version of the manuscript.

DECLARATION OF INTERESTS
K.L. and A.Y. are employees of Insilico Medicine Taiwan. Y.-C.L. is CEO of Insilico Medicine Taiwan. G.J.L.B. and T.R. are co-founders and shareholders of TargTex S.A.

INCLUSION AND DIVERSITY
We worked to ensure ethnic or other types of diversity in the recruitment of human subjects.

Received: May 20, 2021
Revised: August 3, 2021
Accepted: August 23, 2021
Published: September 13, 2021

REFERENCES
1. Garbaccio, R.M., and Parmee, E.R. (2016). The impact of chemical probes in drug discovery: a pharmaceutical industry perspective. Cell Chem. Biol. 23, 10–17.
2. Schürmann, M., Janning, P., Ziegler, S., and Waldmann, H. (2016). Small-molecule target engagement in cells. Cell Chem. Biol. 23, 435–441.
3. Arrowsmith, C.H., Audia, J.E., Austin, C., Baell, J., Bennett, J., Blagg, J., Bountra, C., Brennan, P.E., Brown, P.J., Bunnage, M.E., et al. (2015). The promise and peril of chemical probes. Nat. Chem. 7, 536–541.
4. Dahlin, J.L., Auld, D.S., Rothenaigner, I., Haney, S., Sexton, J.Z., Nissink, J.W.M., Walsh, J., Lee, J.A., Strelow, J.M., Willard, F.S., et al. (2021). Nuisance compounds in cellular assays. Cell Chem. Biol. 28, 356–370.
5. Coan, K.E., Maliky, D.A., Burlingame, A.L., and Shoichet, B.K. (2009). Promiscuous aggregate-based inhibitors promote enzyme unfolding. J. Med. Chem. 52, 2067–2075.
6. Jadhav, A., Ferreira, R.S., Klump, C., Mott, B.T., Austin, C.P., Inglese, J., Thomas, C.J., Maloney, J., Shoichet, B.K., and Simeonov, A. (2010). Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. J. Med. Chem. 53, 37–51.
7. Baell, J.B., and Holloway, G.A. (2010). 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.
8. Feldmann, C., and Bajorath, J. (2021). Machine learning reveals that structural features distinguishing promiscuous and non-promiscuous compounds depend on target combinations. Sci. Rep. 11, 7863.
9. Stork, C., Chen, Y., Siho, M., and Kirchmair, J. (2019). Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. J. Chem. Inf. Model. 59, 1090–1043.
10. Matlock, M.K., Hughes, T.B., Dahlin, J.L., and Swamidass, S.J. (2018). Modeling small-molecule reactivity identifies promiscuous bioactive compounds. J. Chem. Inf. Model. 58, 1483–1500.
11. Reker, D., Bernardes, G.J.L., and Rodrigues, T. (2019). Computational advances in combating colloidal aggregation in drug discovery. Nat. Chem. 11, 402–418.
12. Feng, B.Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B.K., and Austin, C.P. (2007). A high-throughput screen for aggregation-based inhibition in a large compound library. J. Med. Chem. 50, 2385–2390.
13. Bajorath, J. (2021). Evolution of assay interference concepts in drug discovery. Expert Opin. Drug Discov. 16, 719–721.
14. Feng, B.Y., and Shoichet, B.K. (2006). A detergent-based assay for the detection of promiscuous inhibitors. Nat. Protoc. 1, 550–553.
15. Feng, B.Y., Shelat, A., Doman, T.N., Guy, R.K., and Shoichet, B.K. (2005). High-throughput assays for promiscuous inhibitors. Nat. Chem. Biol. 1, 146–148.
16. Walters, W.P., and Murcko, M.A. (2002). Prediction of ‘drug-likeness’. Adv. Drug Deliv. Rev. 54, 255–271.
17. Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L., and Ramsden, N. (1999). Strategic pooling of compounds for high-throughput screening. J. Chem Inf. Comput. Sci. 39, 897–902.
18. Yang, J.J., Utsu, O., Lipinski, C.A., Sklar, L.A., Opres, T.I., and Bologa, C.G. (2016). Badapple: promiscuity patterns from noisy evidence. J. Cheminform. 8, 29.
19. Young, R.J., Green, D.V.S., Luscombe, C.N., and Hill, A.P. (2011). Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. Drug Discov. Today. 16, 822–830.
20. Feldmann, C., Miljkovic, F., Yanchev, D., and Bajorath, J. (2019). Identifying promiscuous compounds with activity against different target classes. Molecules 24, 4185.
21. Lipinski, C.A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today Technol. 1, 357–351.
22. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., and Hopkins, A.L. (2012). Quantifying the chemical beauty of drugs. Nat. Chem. 4, 90–98.
23. Beker, W., Wołos, A., Syzmuk, S., and Grybowski, B.A. (2020). Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks. Nat. Mach. Intell. 2, 457–465.
24. Shultz, M.D. (2019). Two decades under the influence of the rule of five and the changing properties of approved oral drugs. J. Med. Chem. 62, 1701–1714.
25. Yusof, I., and Segall, M.D. (2013). Considering the impact drug-like properties have on the
chance of success. Drug Discov. Today 18, 659–666.
26. Mignani, S., Rodrigues, J., Tomas, H., Jalal, R., Singh, P.P., Majoral, J.-P., and Vatwatwarska, R.A. (2018). Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? Drug Discov. Today 23, 605–615.
27. Stork, C., and Kirchmair, J. (2018). PAINS (Pathway- and Activity-based Indicators of Novel Structures) as computational tools: how much have we simplified? Future Med. Chem. 10, 1533–1535.
28. Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanne, A., Bleicher, K., Daniel, F., Gutknecht, E.-M., Rogers-Evans, M., et al. (2002). Development of a virtual screening method for identification of “frequent hitters” in compound libraries. J. Med. Chem. 45, 137–142.
29. Stork, C., Wagner, J., Friedrich, N.-O., de Bruyn, M., Balogh, B., Ma´ ndity, I.M., et al. (2019). Design and selection of novel C1s inhibitors by in silico and in vitro approaches. Molecules 24, 3641.
30. Slizyk, K., Hajdú, I., Flachner, B., Lórinz, Z., Balcerz, J., Gál, P., Závodszky, P., Pirli, C., Balogh, B., Mándity, I.M., et al. (2019). Design and selection of novel C1s inhibitors by in silico and in vitro approaches. Molecules 24, 3641.
31. Bajorek, A., and Bajorath, J. (2020). Adaptive optimization of chemical reactions with minimal experimental information. Cell Rep. Phys. Sci. 1, 100247.
32. Chuang, K.V., and Keiser, M.J. (2018). Adversarial controls for scientific machine learning. ACS Chem. Biol. 13, 2619–2821.
33. Dusko, V., Grizzou, J., Xuang, W., Hsioni, Z., Long, D.-L., Miras, H.N., and Cronin, L. (2017). Human versus robots in the discovery and characterization of gigantic polyoxometalates. Angew. Chem. Int. Ed. Engl. 56, 10815–10820.
34. Eisses, M.J., Adams, T., Liston, D.E., Low, D.K., Newman, S.F., Kim, J., and Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxemia during surgery. Nat. Biomed. Eng. 2, 749–760.
35. Feng, X., and Bajorek, A. (2020). Identification of compounds that interfere with high-throughput screening assay technologies. ChemMedChem 14, 1795–1802.
36. Hsieh, J.-H., Wang, X.S., Teotico, D., Golbraikh, A., and Tropsha, A. (2008). Development of Ampc beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. J. Comput. Aided Mol. Des. 22, 593–609.
37. Blaschke, T., Miljkovic, F., and Bajorek, J. (2019). Prediction of different classes of promiscuous and nonpromiscuous compounds using machine learning and nearest neighbor analysis. ACS Omega 4, 6883–6890.
38. Feldmann, C., Yonchev, D., Stumpfe, D., and Bajorek, J. (2020). Analysis of biological screening compounds with single- or multi-target activity via diagnostic machine learning. Biomolecules 10, 1605.
39. Ghosh, D., Koch, U., Hadian, K., Sattler, M., and Tetko, I.V. (2018). Luciferase Advisor: high-accuracy model to flag false positive hits in luciferase HTS assays. J. Chem. Inf. Model. 58, 933–942.
40. Irwin, J.J., Duan, D., Torosyan, H., Doak, A.K., Ziebarth, K.T., Strother, T., Tumanian, G., and Shoichet, B.K. (2015). An aggregation advisor for ligand discovery. J. Med. Chem. 58, 7076–7087.
41. Luo, H., Li, Z., Li, X., Mo, X., Ung, C., Li, H., Liu, X., and Chen, Y. (2010). Identification of small molecule aggregators from large compound libraries by support vector machines. J. Comput. Chem. 31, 752–763.
42. Chae, S., and Kirchmair, J. (2018). PAINS (Pathway- and Activity-based Indicators of Novel Structures) as computational tools: how much have we simplified? Future Med. Chem. 10, 1533–1535.
43. Alves, V.M., Capuzzi, S.J., Braga, R.C., Korn, D., Hochuli, J.E., Bowler, K.H., Yagasar, A., Rai, G., Simeonov, A., Muratov, E.N., et al. (2020). SCAM Detective: accurate predictor of small, coloiddalg aggregating molecules. J. Chem. Inf. Model. 60, 4056–4063.
44. Reker, D., Hoyt, E.A., Bernardes, G.J.L., and Rodrigues, T. (2020). Adaptive optimization of chemical reactions with minimal experimental information. Cell Rep. Phys. Sci. 1, 100247.
45. Engkvist, O. (2019). Identification of Ampc beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. J. Comput. Aided Mol. Des. 22, 593–609.
46. Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., et al. (2020). Molecular Sets (MOSES): a benchmarking platform for molecular generation models. Front. Pharmacol. 11, 56644.