CHEMINFORMATICS ANALYSIS AND COMPUTATIONAL MODELING OF DETERGENT-SENSITIVE AGGREGATION

Kyle Hopkins Bowler

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Division of Chemical Biology and Medicinal Chemistry in the Eshelman School of Pharmacy (Pharmaceutical Sciences).

Chapel Hill
2019

Approved by:

Michael B. Jarstfer

Tim Willson

Alexander Tropsha

ABSTRACT

Kyle Hopkins Bowler: Cheminformatics Analysis and Computational Modeling
of Detergent-Sensitive Aggregation
(Under the direction of Alexander Tropsha)

Small molecule aggregates cause detergent reversible protein sequestration and are the most prevalent source of nonspecific activity in biochemical screening assays. Large volumes of publicly available dose-response screens performed in the presence or absence of detergent have enabled cheminformatics analysis into chemical aggregation which reinforces prior notions that aggregation is prevalent and context dependent. We report the development of random forest classifiers trained on screens of β-lactamase or cruzain targets under well-defined assay conditions which distinguish putative aggregators and non-aggregators with balanced accuracies as high as 78%. These models overcome limitations of existing computational predictors related to programmatic errors, blurred modeling endpoints, and poor external predictivity. Model interpretation indicated that polarity, aliphaticity, and weight are significantly correlated with aggregation propensity, although these features alone estimate behavior with under 70% accuracy. Our curated datasets and validated models will help identify potential aggregators and reduce resource waste during drug discovery and optimization.

To my family and friends, whose support made this work possible.
Thank you to my parents, Kimberly and Rodney, my sister Emily,
and my closest friends, Taylor and Phoebe.

ACKNOWLEDGEMENTS

I gratefully thank Stephen J. Capuzzi, Vinicius Alves, Eugene Muratov, and Alexander Tropsha, my scientific mentors who significantly contributed to this work. I also thank Alexey Zakharov from the Computer-Aided Drug Design Group in the Center for Cancer Research at NCATS for interpretations of experimental data related to mechanisms of assay interference. I especially appreciate the significant efforts from Brian Shoichet *et al.* at the University of California, San Francisco in partnership with NCGC in generating and sharing the datasets used for this work. Additionally, I am grateful for my committee members Michael B. Jarstfer and Tim Willson for their contributions and constructive criticisms. I appreciate feedback from Jonathan Baell from the Monash Institute of Pharmaceutical Sciences for informing the aggregation hill-slope threshold and from Kenneth Pearce regarding the target audience and functionality of the predictors. I acknowledge the KNIME community and RDKit for key contributions in cheminformatics analytics tools and the Chemical Computing Group, Kode Chemoinformatics, Molinspiration, ISIDA, and ChemAxon for their software licenses or trials. Finally, I thank my lab mates within the Laboratory for Molecular Modeling and peers in my cohort for their support, advice, and camaraderie.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATIONS x

CHAPTER 1: CHEMINFORMATICS ANALYSIS AND COMPUTATIONAL MODELING OF DETERGENT-SENSITIVE AGGREGATION 1

1. INTRODUCTION 1

2. MATERIALS AND METHODS 7

2.1 Data Collection 7

2.2 Concentration Dependence 8

2.3 Data Classification 8

2.4 Data Curation 9

2.5 Context Dependence 11

2.6 External Software 11

2.6.1 The Aggregator Advisor 11

2.6.2 Hit Dexter 13

2.7 Molecular Descriptor Calculation 13

2.8 Feature Processing and Training Set Balancing 14

2.9 Modelability Index 15

2.10 QSPR Model Generation and Validation 16

2.11 Interpretation of QSPR Models 18

3. RESULTS 20
3.1 Concentration Dependence 20
3.2 Dataset Concordance 20
3.3 Context Dependence 21
3.4 Aggregator Advisor Analyses 22
3.4.1 Classification Agreement 22
3.4.2 Programmatic Evaluation 24
3.5 Hit Dexter Analysis 25
3.6 MODI and QSPR Model Performance 26
3.7 Interpretation of QSPR Models 28
4. DISCUSSION AND CONCLUSIONS 35
5. DISSEMINATION 46
APPENDIX 48
REFERENCES 52

LIST OF TABLES

Table 1 - Aggregator concentration dependence for PubChem datasets 20

Table 2 - Intra-dataset concordance analysis 20

Table 3 - β-lactamase inter-dataset concordance analysis 21

Table 4 - Inter-target concordance analysis 21

Table 5 - Aggregator sources within the Advisor 22

Table 6 - Rogues' Gallery Aggregators categorized by Advisor logic 24

Table 7 – Hit Dexter error/warning frequency 26

Table 8 - Training set MODI calculations 26

Table 9 - Predictive statistics of β-lactamase models 27

Table 10 - Predictive statistics of cruzain models 27

Table 11 - Model-prioritized Dragon7 features 29

Table 12 - Model-prioritized ISIDA features 32

Table 13 - Predictive statistics for RDKit β-lactamase models with optimized thresholds 34

Table 14 - Predictive statistics for the refined, hybrid-feature β-lactamase RF model 35

LIST OF FIGURES

Figure 1 - Overall study design 6

Figure 2 - Per-class dataset size throughout data curation 9

Figure 3 - General modeling workflow 16

Figure 4 - Overlapping compounds tested between targets 21

Figure 5 - Overlapping compounds used in the Advisor and tested in either target 23

Figure 6 – Overlapping compound class agreement between target sets and the Advisor 23

Figure 7 - Inconsistent results for an Aggregator tested in two β-lactamase assays 25

Figure 8 - Per-class Hit Dexter promiscuity probability distributions 25

Figure 9 - Per-class distributions for select Dragon7 model-prioritized features 30

Figure 10 - Per-class distributions for MLOGP 31

Figure 11 - Per-class distributions for C-H bond count 32

Figure 12 - Per-class distributions for select RDKit features 34

LIST OF ABBREVIATIONS

5FCV	5-fold external cross validation
AC_{50}	Concentration required to elicit 50% of an active response
AD	Applicability domain
Agg	Aggregator
AID	PubChem assay identifier
CAC	Critical aggregating concentration
CID	PubChem compound identifier
csv	Comma-separated values file
Da	Daltons
DTT	Dithiothreitol
FN	False negative
FP	False positive
HTS	High-throughput screening
IC_{50}	Concentration required to elicit 50% of an inhibitory response
IUPAC	International union of pure and applied chemistry
LogP	Octanol-water partition coefficient
miLogP	LogP value calculated by Molinspiration
MLOGP	LogP value calculated by Dragon7
MODI	Modelability index
MW	Molecular weight
NCATS	National center for advancing translational sciences at the NIH
NCGC	NCATS chemical genomics center

nM	Nanomolar
nm	Nanometers
NN	Nearest neighbor
Non-Agg	Non-aggregator
NPV	Negative predictivity
pH	Negative log of the concentration of Hydrogen atoms
PPV	Positive predictivity
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
RF	Random forest
SDF	Structure data format
SE	Sensitivity
SIR	Structure-interference relationship
SlogP	LogP value calculated by RDKit
SMILES	Simplified molecular input line entry system
SP	Specificity
TC	Tanimoto similarity coefficient
TN	True negative
TP	True positive
VS	Virtual screening
v/v	Volume per volume concentration
y-rand	Y-randomization
μM	Micromolar

CHAPTER 1: CHEMINFORMATICS ANALYSIS AND COMPUTATIONAL MODELING OF DETERGENT-SENSITIVE AGGREGATION

1. Introduction

The premier method for initial hit identification in the modern pharmaceutical pipeline is High-Throughput Screening (HTS). This approach often involves massive biochemical assaying of biological target(s) against organic small molecule libraries to identify potentially active chemical leads.[1–3] Although initial HTS campaigns have reported apparent hit rates as high as 5% for certain targeted libraries, subsequent orthogonal testing has revealed that the frequency of specific, biologically active compounds is generally fewer than 0.1%.[4] False-positive compounds are a prevalent liability in HTS libraries and confound the interpretation of assay readouts by interfering through distinct mechanisms with assay systems or detection technology rather than directly interacting with desired biological targets.[5,6] Regrettably, significant resources are wasted on the progression and ultimate failure of false-positive assay interference compounds in the drug development pipeline.[7–9]

The most common source of biochemical assay interference artifacts is through the formation of colloidal aggregates.[4,10] For some HTS campaigns, at least 90% of primary screening hits have been found to exhibit promiscuous activity as aggregators.[3,11,12] Aggregation is a well-described physical phenomenon that involves the accumulation and precipitation of small molecules in bubble-like colloids.[10,12,13] Aggregate precipitation can significantly disrupt *in vitro* assays by sequestering and partially unfolding target proteins in solution through reversible and high affinity adsorption to the surface of colloidal particles.[14–17] This specific mechanism can elicit

assay artifacts and false-positive readouts in the context of biochemical HTS by presenting misleading data which suggests target inhibition but in reality is evidence of nonspecific activity. While aggregation-mediated interference has been implicated in other biological assay formats, such as cell-based phenotypic screens and *in-vivo* assays, the mechanism and pervasiveness in these contexts are less well characterized.[12]

The affinities of specific proteins to adsorb to specific small molecule aggregates are highly variable[3,10,18] and although sequestration and perturbation of tertiary structure most commonly results in the reduction of apparent protein activity, the perceived activity of some proteins may not be sensitive to, or may be enhanced by, aggregate adsorption.[3] Additionally, colloid formation is context dependent and sensitive to changes in assay buffer system, pH, temperature, and mixing protocols[12], and the structure of the aggregating compound also impacts the Critical Aggregating Concentration (CAC), which is commonly in the low micromolar range and is characterized by a steep hill-slope curve.[12,19]

Colloids tend to reach mid-femtomolar concentrations and are imperceptible to the naked eye and non-trivial to detect.[10,13] Therefore, experimental assay considerations and orthogonal controls which identify or triage aggregating compounds have been developed.[12] The most common and effective approach for mitigating aggregation in biochemical assays has been the addition of detergents into assay reaction mixtures which disrupt colloids and raise the CAC.[14,18,20] For example, the addition of 0.01% (v/v) non-ionic detergent Triton X-100 effectively prevents colloid formation for most small molecules, and the calculation of detergent-sensitivity is a powerful proxy of the physical phenomenon of aggregation.[21] However, some colloids are not detergent sensitive[21], and not all assay formats can accommodate detergents due to potential interference with the target, reagents, or instrumentation.[12] Other assay alterations and *post hoc*

methods which account for aggregators in biochemical screens also have general drawbacks. For example, these techniques may be expensive, low- or mid-throughput, require specialized instrumentation or excess protein and compound concentrations, or augment assay conditions in ways that may not be compatible.[12] These drawbacks motivate the development of supplemental cheminformatics approaches which are robust, inexpensive, and enable rapid prospective screening of HTS libraries.

Significant efforts have been made to gain insight on the Structure-Interference Relationship (SIR) of detergent-sensitive aggregators. For example, substantial dose-series HTS against AmpC β-lactamase and cruzain targets have been performed both with and without the addition of detergent.[11,12,22,23] PubChem assay IDs 485341, 485294, 585, and 584 for β-lactamase and 1476, 1478 for cruzain detail several hundred thousand datapoints for which an organic small molecule was tested at multiple concentrations, resulting in dose response curves which provide insight into compounds which may have activities sensitive to the addition of 0.01% Triton X-100 detergent. The publishing of massive HTS datasets such as these has empowered the development of cheminformatics approaches which aim to accurately predict future aggregating compounds.[24–26]

Quantitative Structure-Property Relationship (QSPR) models have been developed towards the specific endpoint of aggregation. QSPR and Quantitative Structure-Activity Relationship (QSAR) modeling are well established techniques invaluable to the field of medicinal chemistry which attempt to predict experimental endpoints for compounds using features derivable from chemical structure.[27] The routine of building QSAR/QSPR models involves the collection of suitably diverse, adequately sized datasets of experimental results paired with tested chemical structures, the standardization of structures and curation of data, the balancing of datasets to achieve equal data distribution across categorical or continuous endpoints, the calculation of

structural descriptors, the partitioning of data into model training sets and external testing sets, the development of statistical models which are commonly machine learning-based, and most critically, the validation of trained models through prediction of withheld external test sets within a well-defined applicability domain and the execution of y-randomized models which detect dataset overfitting.[27] Validated models should be mechanistically interpreted to gain insight into the endpoint being modeled[28], and their primary application is in the Virtual Screening (VS) of additional chemical structure libraries to make fast and inexpensive predictions of activities and properties of compounds to be prospectively tested experimentally.[27]

Several groups have estimated aggregation or general promiscuity through statistical, QSPR-based approaches.[23,24,26,29,30] For example, OpenEye software available at https://www.eyesopen.com/ features a recursive partitioning classifier trained on 47 aggregators and 64 non-aggregators for which 260 descriptors were calculated.[24] Later, a support vector machines aggregator classifier was trained on 1,319 aggregators and 128,325 non-aggregators based on classifications from the β-lactamase HTS data published by the Shoichet Laboratory at the time.[26] Recently, Hit Dexter has been released as a statistical approach for predicting frequent false-positive compounds in general.[29,30]

While these approaches have demonstrated impressive classification performance on historical datasets, previously-developed QSPR models which predict aggregation are also at risk to present high false-negative and false-positive classification rates.[23–26] Models tested externally were found by Feng *et al.*[23] and Shoichet *el al.*[25] to have low predictive extensibility, likely as a result of the limited diversity of the training sets with several hundred aggregating compounds. Additionally, the Hit Dexter 1.0 tool[29] which predicts highly promiscuous compounds in general is not mechanistically specific to the nuanced endpoint of aggregation and, in our analysis, is too

broad to accurately estimate aggregation, although the newer 2.0 version[30] has not yet been evaluated.

The most popular tool for computationally flagging aggregating compounds is the Aggregator Advisor (http://advisor.bkslab.org/search/).[12,25] This online application hosts over twelve thousand compounds deemed to display detergent-sensitive target inhibition from eighteen literature sources since 2015. Specifically, molecules with an IC_{50} less than 100 μM which was diminished or eliminated by the addition of detergent were classified as aggregators and were placed in their "Rogues' Gallery" reference database, which can also be found as a catalog on the ZINC15 website[31] at https://zinc15.docking.org/catalogs/aggregators/. Users may query the tool with chemical SMILES and the Advisor employs a proprietary algorithm through ChemAxon's AxonPath fingerprints to calculate the Tanimoto Similarity Coefficient (TC)[32] of the query to all aggregators in the Rogues' Gallery. Additionally, the tool approximates chemical lipophilicity by calculating the logP value of the query through the Molinspiration software at https://www.molinspiration.com/, which is a powerful indicator of chemical solubility and drug-likeness. If the TC to a "Rogue" is greater than 80% or the miLogP value is greater than 3.0, the user is informed that this compound should be investigated as an aggregator and that controls are always advised.

However, in our analysis of the Aggregator Advisor[25], theoretical and programmatic issues surfaced which weaken the tool's capability of identifying future aggregators. First, the quantitative threshold which describes detergent-sensitive aggregation is not consistent between the 18 orthogonal sources used to develop the "Rogues' Gallery". This, combined with the amalgamation of data from diverse assay experimental conditions, blurs the specific modeling endpoint and violates the concept that aggregation is a context dependent phenomenon.

Additionally, the TC search algorithm employed by the Advisor cannot extend to the prediction of new chemotypes which may aggregate but are not related to the static reference dataset. Finally, we identified programmatic errors with the Advisor where 266 compounds from the Rogues' Gallery were found through query to have neither miLogP greater than 3.0 nor TC greater than 80% to any compound within the Rogues' Gallery.

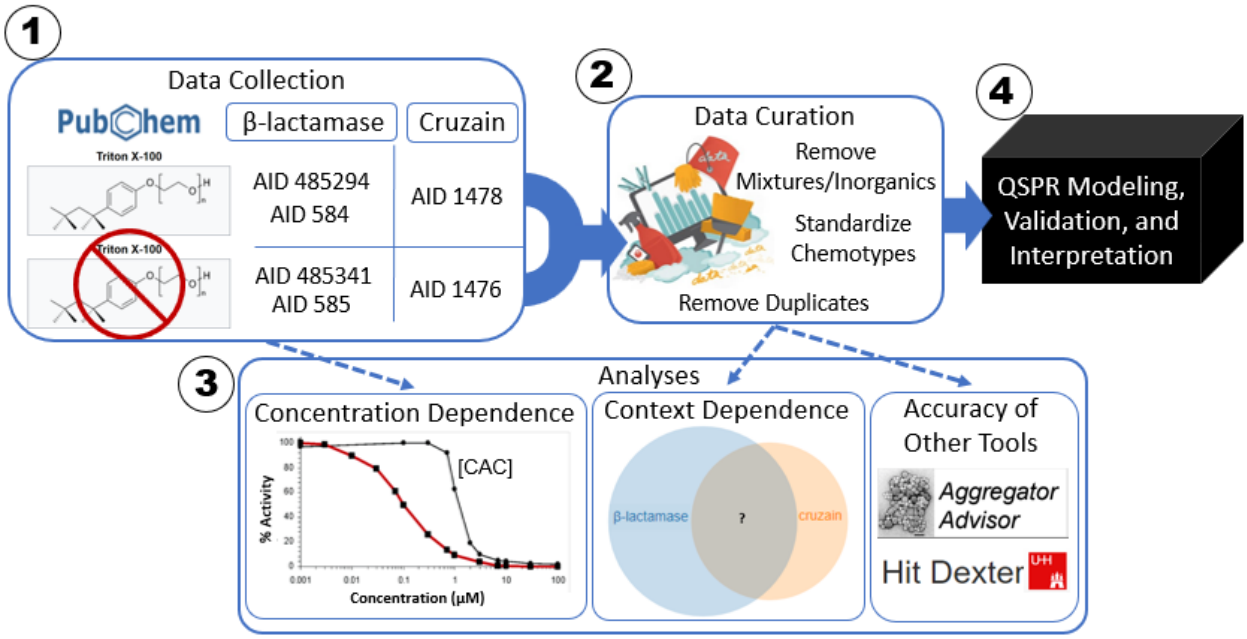


Figure 1 - Overall study design

The present study consists of (1) data collection, (2) data curation, (3) analyses probing concentration and context dependence and historical aggregation predictors, and (4) QSPR modeling, validation, and interpretation. This figure uses adapted images from the following cited references, including https://pubchem.ncbi.nlm.nih.gov/.[25,27,29,33]

The wealth of public HTS data on detergent-sensitive, small molecule aggregators has enabled unprecedented modeling for the prediction of aggregation. Here we report the development and rigorous validation of reliable and extensible QSPR models of detergent-sensitive aggregation for separate β-lactamase and cruzain targets with defined assay conditions and classification thresholds (**Figure 1**).

2. Materials and Methods

2.1 Data Collection

HTS campaigns corresponding to a pair of PubChem assay IDs (AIDs) in which an organic small molecule library was screened both with and without the addition of 0.01% Triton X-100 detergent were downloaded from https://pubchem.ncbi.nlm.nih.gov/bioassay. This includes two separate campaigns against AmpC β-lactamase tested under identical assay conditions on different chemical libraries and at variable concentration intervals, corresponding to AID pairs 485341/485294 and 585/584.[23] These assays were performed in 20 mM phosphate buffer at pH 7.0 using 5.3 nM AmpC β-lactamase protein in 1536-well Greiner black clear bottom plates at room temperature and activity was detected every twenty seconds for four minutes on a ViewLux (Perkin-Elmer) High-throughput CCD imager using 480 nm absorbance. Compounds were tested at 0.457 μM, 2.290 μM, 11.40 μM, and 57.10 μM concentrations for AID pair 485341/48529 and at 1.8 nM, 9.2 nM, 0.046 μM, 0.23 μM, 1.15 μM, 5.75 μM, and 28.73 μM concentrations for AID pair 585/584. A single campaign was extracted against the cysteine protease cruzain, corresponding to the AID pair 1476/1478.[11,22] These assays were performed in 100 mM sodium acetate at pH 5.5 with 5 mM DTT using 1.5 nM cruzain protein in 1546-well Greiner black solid bottom plates at room temperature and activity was detected four times every thirty seconds for two minutes on ViewLux High-throughput CCD imager (Perkin-Elmer) using 360 nm excitation and 450 nm emission fluorescence. AID pairs were joined on PubChem compound ID (CID) chemical identifiers, resulting in 320,095 datapoints with dose-response data both with and without detergent for the 485341/485294 AID pair, 70,660 datapoints for the 585/584 AID pair, and 197,809 datapoints for the 1476/1478 AID pair.

2.2 Concentration Dependence

Datapoints for the two HTS campaigns against β-lactamase were concatenated and -$Log_{10}(AC_{50})$ (pAC_{50}) values from the detergent-free assays were binned according to concentration range separately for each target. The percentage of the total number of active datapoints for each range was calculated. In this context, active data refers to datapoints which have real, continuous pAC_{50} values reported in the detergent-free assay raw data.

2.3 Data Classification

We assign a binary classification to compounds based on aggregators displaying sharp hill-slopes and activity which is completely abolished by the addition of detergent.[3,12,14,33] Several HTS groups such as Southall *et al.* have adopted a dose-response curve classification ontology which labels completely inert curves as class 4.0.[34] Collected data tested at serial concentrations were accompanied by fitted values approximating hill-slope, pAC_{50}, and curve class. In this study, compounds possessing a hill-slope greater than 1.5 with a real pAC_{50} value which are not curve class 4.0 without detergent, but which are curve class 4.0 and have no pAC_{50} value reported with detergent are labeled "putative aggregators". Compounds with no reported pAC_{50} value which are curve class 4.0 with and without detergent are labeled "non-aggregators". This classification resulted in 26,212 putative aggregators and 239,359 non-aggregators for AID pairs 485341/485294, 13,825 putative aggregators and 38,207 non-aggregators for AID pairs 585/584, and 25,360 putative aggregators and 168,849 non-aggregators for AID pairs 1476/1478. A third classification which was not used for modeling was defined. Compounds possessing a hill-slope less than or equal to 1.5 with a real pAC_{50} value which are not curve class 4.0 with and without detergent were labeled "detergent-resistant actives". This resulted in 55 detergent-resistant actives

for AID pairs 585341/485294, 33 for AID pairs 585/584, and 155 for AID pairs 1476/1478. Compounds which do not satisfy any classification were removed from the dataset.

2.4 Data Curation

Classified data were curated per best practices of molecular modeling to cleanse chemical data which may be incomplete, inaccurate, imprecise, incompatible, or irreproducible (**Figure 2**).[35–37]

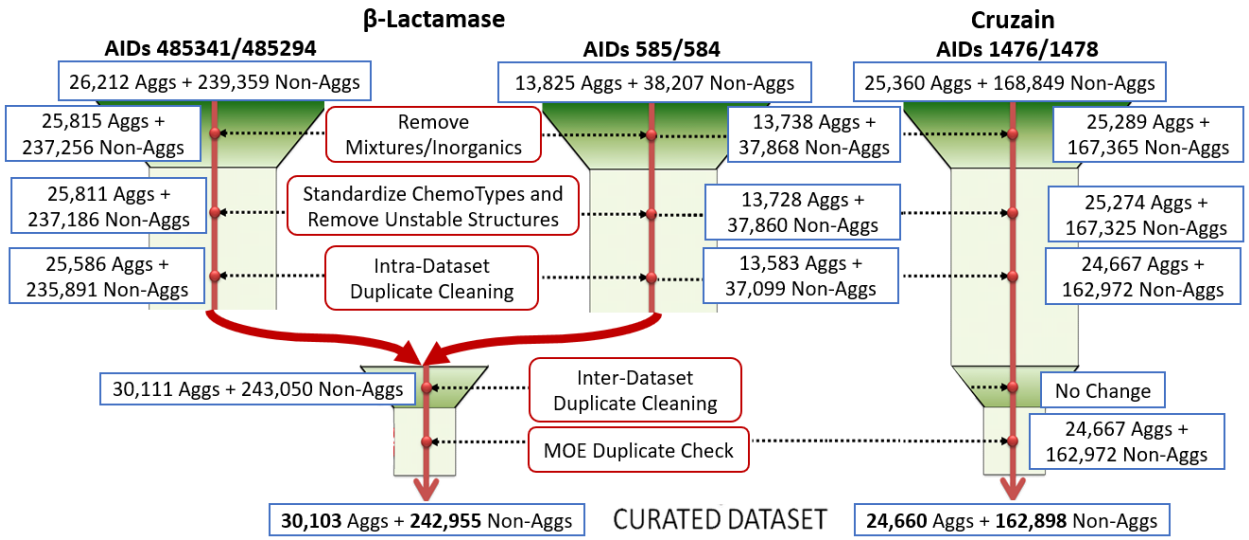


Figure 2 - Per-class dataset size throughout data curation
The number of putative aggregator (Aggs) and non-aggregator (Non-Aggs) compounds at each step in the curation protocol. Ordered steps include (1) removal of mixtures and inorganics, (2) standardization of chemotypes and removal of unstable resulting structures, (3) intra-dataset duplicate cleaning and inter-dataset duplicate cleaning for the β-lactamase dataset, and (4) a duplicate check using MOE, resulting in final curated datasets. Curation workflow and image adapted from prior works.[37]

SDF chemical structures were obtained from the PubChem Download Service at https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi and were joined to classified compounds on CID values. Mixtures and inorganic compounds were removed by rendering compounds as RDKit objects, verifying electronic valence, separating the two largest components which are not covalently bound, removing inorganics with the largest component missing carbon atoms, removing mixtures with the second largest component possessing carbon atoms, and removing

salts and solvents by maintaining only the largest component. Chemical chemotypes were standardized using ChemAxon's "Standardizer" software available at https://chemaxon.com/products/chemical-structure-representation-toolkit. The following chemical cleaning operations were applied to each chemical structure in the given sequence: isotope annotations were stripped, explicit hydrogens were removed, rings were dearomatized, stereocenter specifications were removed, compounds were neutralized to have no explicit charges, all absolute stereospecificity was cleared, wedges were cleaned, compounds were mesomerized, tautomerized, and aromatized, explicit hydrogens were added, specific chemotypes were standardized such as nitro- and aromatic-groups, structures were transformed to neutral chemotypes, and were cleaned in 2D. Structures were dearomatized, rendered as RDKit objects, and converted to IUPAC InChiKey to assure stability following standardization. Assay datapoints were aggregated on InChiKey and percent intra-dataset concordance was calculated per HTS campaign by dividing the number of compounds with more than one assay datapoint which unanimously agree in aggregator classification by the total number of compounds with more than one assay datapoint. The number of assay datapoints for each compound was reduced to one by concatenating all compounds with only one assay datapoint with the first entry from compounds in duplicate sets which unanimously agree. All entries from duplicate assay datapoints which disagreed in classification were removed completely from the dataset to increase data reliability. Similarly, inter-dataset duplicate analysis was performed on the data from the two β-lactamase HTS campaigns. Compounds from the 485341/485294 AID pair and 585/584 AID pair were joined on InChiKey and the percent inter-dataset concordance was calculated by dividing the number of compounds which exist in both sets and which agree in aggregator classification by the total number of compounds which exist in both sets. To merge the two datasets, compounds which were unique to either HTS campaign were

concatenated with the compounds which agreed between the two campaigns. This process resulted in the concatenation of 22,774 putative aggregators, 206,647 non-aggregators, and 52 detergent resistant actives unique from AID pair 485341/485294 with 5,221 putative aggregators, 13,405 non-aggregators, and 25 detergent resistant actives unique from AID pair 585/584 and finally with 2,116 putative aggregators, 22,998 non-aggregators, and 3 detergent resistant actives which were agreed on between both sets. MOE software was used to remove residual duplicates which share a tautomer based on SDF with a molecule previously seen in the dataset in a sequential fashion. Prior to filtering, the putative aggregation, non-aggregation, and detergent-resistant active compounds were concatenated per target to ensure that if duplicate entries exist, the detergent-resistant actives are those to be removed. The 68 curated detergent-resistant actives for β-lactamase and 151 for cruzain were split out of the data and separately saved, while the remaining datasets were rendered as RDKit objects once more to verify chemical stability. Curated datasets were found to contain a small number of cis- and trans- E-Z stereoisomers.

2.5 Context Dependence

The two curated target datasets with putative aggregators, non-aggregators, and detergent-resistant actives were joined based on InChiKey identifiers. Overlapping compounds were aggregated based on their binary classification in both target datasets. For this analysis, detergent-resistant actives were also assigned to the non-aggregator class.

2.6 External Software

2.6.1 The Aggregator Advisor

The 12,641 compounds within the Advisor's Rogues' Gallery reference database were acquired at http://advisor.bkslab.org/rawdata/aggpage.txt. These SMILES were rendered as RDKit objects and converted through RDKit to SDF, resulting in the loss of two unreadable

structures. These structures were curated as previously described. This process resulted in the filtration of 30 aggregators possessing mixtures or inorganics, 5 aggregators with unstable structures following chemotype standardization, and 264 aggregators which were extra duplicate values. The curated Rogues' Gallery was separately joined on InChiKey to the curated β-lactamase and cruzain datasets. For this comparison the detergent-resistant actives were not included.

The reference Rogues' Gallery dataset was aggregated on chemical SMILES to isolate 12,608 unique strings from the eighteen orthogonal sources enumerated at http://advisor.bkslab.org/rawdata/aggref.txt. These sources were investigated for the consistency of their quantitative definition of aggregation. These unaltered SMILES were run through the published command line python program "simi.py" provided at http://advisor.bkslab.org/faq/, which was slightly modified to write results to a csv file and return the query molecule SMILES, matched aggregator SMILES, Tanimoto Similarity Coefficient (TC), and compound identifier each time a query matched any given SMILES in the reference dataset with a TC of at least 0.70. These 12,608 SMILES were run through the proprietary Molinspiration software provided at http://www.molinspiration.com/cgi-bin/properties to calculate the miLogP property using the mib batch molecule processing v2017.01 program. Two SMILES were flagged as having strange valence at oxygen by the tool and were removed from the dataset. Compounds were characterized as aggregators through the same logic published in the Aggregator Advisor tool: they either had a miLogP value greater than 3.0 or a TC value to a matched compound in the Advisor tool greater than 0.8. A handful of compound SMILES were copied and pasted directly from the reference dataset into the Aggregator Advisor online application at http://advisor.bkslab.org/ and queries were run at each available affinity range of either < 0.1 μM, 0.1 – 10 μM, or > 10 μM to test

whether the online application behaved the same as the batch python script (**Supplemental Material 1**).

Several compounds were isolated which had less than 0.70 TC to any compound in the Aggregator Advisor tool and a miLogP less than 3.0. "COC(\O)=c2/sc1cc(C)ccc1nc2C" with CID 713501 was selected for examination of experimental data. This compound was queried through PubChem at https://pubchem.ncbi.nlm.nih.gov/search/ and was evaluated for its activity in HTS assays.

2.6.2 Hit Dexter

Curated and classified β-lactamase and cruzain detergent-sensitive aggregator datasets were rendered as RDKit objects, converted to SMILES format, and screened through the Hit Dexter online tool at https://hitdexter.zbh.uni-hamburg.de/ (version 1.0).[29] The computationally expensive Tanimoto Similarity-based applicability domain was not calculated due to the massive sizes of the datasets. Hit Dexter-predicted "High" and "Moderate or High Promiscuity Probabilities" and the frequency of specific predictive warnings and errors were evaluated per-class.

2.7 Molecular Descriptor Calculation

Molecules were described using two orthogonal feature generation techniques. An amalgam of 4,529 2D whole-molecule physicochemical features were calculated without rounding atom coordinates using the Dragon7 software available from Kode at https://chm.kode-solutions.net/products_dragonknime.php.[38] Additionally, the ISIDA Substructural Molecular Fragments Method[39] (Version 5.8.11, Build 14.01.2018) was employed to calculate the number of occurrences of all Substructural Molecular Fragments within each given dataset. Default fragment calculation settings were used, with the exception that all paths were calculated for atom bond

lengths between 2 and 7. The resulting descriptor files were partitioned to separate the descriptor matrix from the fragment indices, and indexed column headers were replaced by their associated Atom and Bonds string. This resulted in 30,585 unique features for β-lactamase and 29,215 unique features for cruzain. Dragon7 and ISIDA features were also calculated for the detergent-resistant actives.

2.8 Feature Processing and Training Set Balancing

The larger non-aggregator class was subset to match the size of the putative aggregator class for both the β-lactamase and cruzain training sets. Dragon7 feature sets were processed to remove columns with incorrect values, low variance, or high correlation to other features. Dragon7 assigns "-999" values as a result of computational errors, therefore columns with any instance of "-999" or missing values were removed. Columns with low variance typically do not effectively discriminate compound classification, therefore features with less than 0.01 variance were removed. Correlating columns may provide redundant information and were also removed. Towards this, the Pearson's product-moment coefficient was calculated for each pair of features, resulting in the generation of a square matrix of correlation values between -1 and 1. A column was considered to correlate to another if the correlation coefficient was at least +/-0.90. For each correlated family of columns, the count of correlated columns each column has with others in the dataset was determined and used to maintain only the column with the least number of correlated partners. This process was repeated until no more correlating columns existed in the dataset. Finally, all features are normalized to the 0.0 and 1.0 range through linear transformation to ensure that they each carry the same magnitude of weight in the similarity search procedure. This resulted in 861 processed Dragon7 features for β-lactamase and 854 features for cruzain. Datasets were partitioned on their binary classification and a similarity search function was performed mapping

non-aggregating compounds against the putative aggregator set. The Euclidean distance in normalized Dragon7 feature space between each non-aggregator and its nearest neighbor (NN) in the putative aggregator set was calculated. To select for a set of non-aggregators matching the size of the smaller putative aggregator class, non-aggregators were sorted based on their distance to their NN in the putative aggregator dataset. The 50% nearest and 25% furthest non-aggregators to the putative aggregator class were concatenated with an additional 25% random remaining non-aggregators.

Compounds selected through balancing were extracted from the original, unprocessed Dragon7 and ISIDA feature sets and were processed once more to remove columns with "-999" or missing values, columns with less than 0.01 variance, and columns with at least +/-0.90 (Dragon7) or +/-0.60 (ISIDA) correlation to other columns in the dataset. This procedure resulted in curated, balanced, feature processed training datasets of 30,103 compounds per-class for β-lactamase with 890 Dragon7 features and 1,085 ISIDA features and 24,660 compounds per-class for cruzain with 906 Dragon7 features and 1,162 ISIDA features. This left 212,852 and 138,238 non-aggregators externally withheld for the β-lactamase and cruzain test sets, respectively. Compounds from the non-aggregator class which were withheld from the training set due to undersampling were maintained and used later as an external screening set for model validation, discussed below. There exist 42 cis/trans pairs of E-Z stereoisomers in the balanced β-lactamase dataset and 23 pairs in the balanced cruzain dataset which are not differentiated using Dragon7 and ISIDA features.

2.9 Modelability Index

The Modelability Index (MODI) is an *a priori* estimate of the ability to produce predictive models from a dataset based on 1-Nearest Neighbor (NN) learning approach without feature selection.[40] Balanced, feature-processed training sets were normalized to the 0.0 and 1.0 range

through linear transformation and the NN in Euclidean distance in both Dragon7 and ISIDA feature spaces was determined for each compound. The MODI[40] was calculated by taking the average between the ratios of the number of compounds with NNs of the same class divided by the number of total compounds in that class for both the putative aggregator and non-aggregator classes.

2.10 QSPR Model Generation and Validation

QSPR models were built and rigorously validated by several external validation methodologies and controls including cross validation, external screening of withheld sets, and y-randomized modeling, as is best practice for QSAR/QSPR (**Figure 3**).[27,28,41,42]

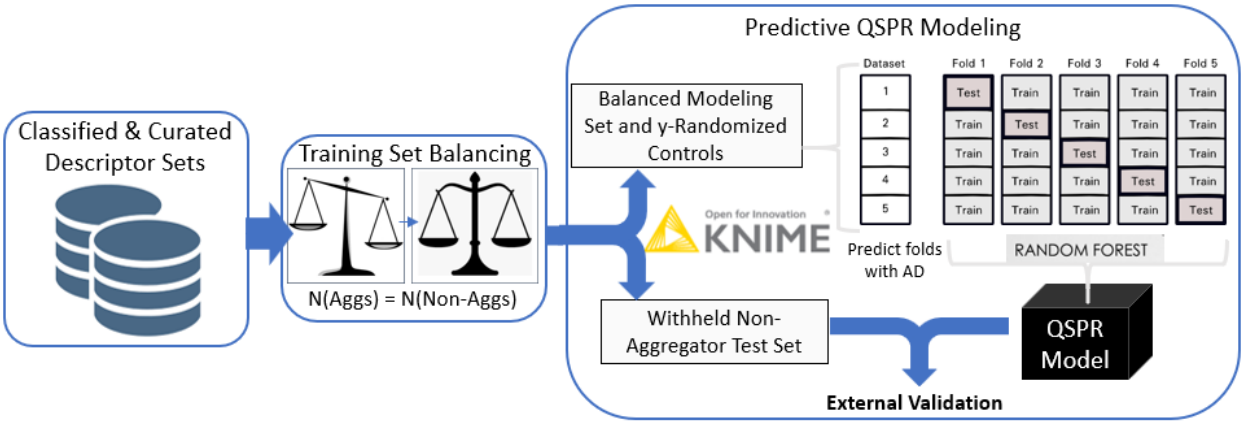


Figure 3 - General modeling workflow

Descriptor sets were processed, and binary classifications were balanced via stochastic and distance-based selection from the non-aggregator class. Balanced sets were modeled using a random forest statistical learning method in five iterations of external cross validation (5FCV) and were implemented in the KNIME Cheminformatics Environment. Y-randomized controls were performed in parallel. Model ensembles were externally validated using non-aggregating compounds which were not selected for the balanced modeling set. An Applicability Domain (AD) was applied based on agreeing consensus predictions from Dragon7 and ISIDA models.

Balanced, feature-processed, and non-normalized training set data were partitioned five separate times to iteratively equip 80% of the compounds for modeling and withhold 20% of the compounds for external validation. Compounds were partitioned based on a stratified sampling approach to ensure an equivalent proportion of putative aggregators and non-aggregators in the

training and test sets. In this way, each compound served exactly four times in the training of new models and exactly once as an external test case. Binary random forest[43,44] modeling was performed using an ensemble of 1,000 trees. Trees were decorrelated by randomly bootstrapping compound instances used in modeling with replacement and selecting a random sample of root(N)-many features for each tree, where N is the total number of features available. Trees were configured to evaluate features on classification accuracy at the median value and to use the information gain ratio as the split criterion. The depth that each tree was permitted to grow was not restrained for Dragon7 models but was limited to a depth of 1,000 for ISIDA models due to computational demand. Predictions were performed using a continuous probability based on the fraction of trees in the forest which vote to classify a given compound in the putative aggregator class. During 5FCV, a single forest of trees was used to screen compounds in the external fold, but for external test sets withheld from modeling the average putative aggregator-probability was calculated for the ensemble of all five forests. Compounds with a putative aggregator-probability greater than 0.5 were predicted to be putative aggregators, and otherwise were predicted to be non-aggregators. To obtain consensus predictions, binary predictions for each compound between the Dragon7 and ISIDA models were joined on InChiKey. Compounds with disagreeing binary class predictions were considered outside of the models' Applicability Domain (AD) and were not screened. Chemical coverage was calculated by dividing the number of compounds within the AD with the total number of screened compounds.

A binary scorer was applied to compounds within the model AD which calculates the number of true positives (TP or putative aggregators correctly predicted), true negatives (TN or non-aggregators correctly predicted), false positives (FP or non-aggregators incorrectly predicted), and false negatives (FN or putative aggregators incorrectly predicted). These values were used to

calculate the following predictivity statistics: sensitivity (SE or TP / (TP + FN)) which estimates the accuracy of the putative aggregator classifications, specificity (SP or TN / TN + FP)) which estimates the accuracy of the non-aggregator classifications, positive predictivity (PPV or TP / (TP + FP)) which estimates the reliability of the putative aggregator predictions, negative predictivity (NPV or TN / (TN + FN)) which estimates the reliability of the non-aggregator predictions, and balanced accuracy ((SE + SP) / 2). Binary classifications for compounds in training datasets were randomly shuffled to maintain class distribution but perturb the link between these classes and their features. Y-randomized modeling was performed using identical configuration settings as the non-randomized modeling described above and predictivity statistics for these models were calculated. Consensus models were evaluated on two external testing sets not included within the training dataset. The non-aggregators withheld during balancing and the detergent-resistant active sets which were curated and processed alongside the training set compounds were screened. In this context, the detergent-resistant actives were considered non-aggregators. Only coverage and SP metrics were calculated for these sets because they were exclusively composed of non-aggregating compounds.

Finally, the predictivity of the Dragon7 and ISIDA models were evaluated independently for accuracy in 5FCV, y-randomization, and external validation, although no AD was applied.

2.11 Interpretation of QSPR Models

Model attribute statistics were used to calculate the percent each feature was chosen at the first split point in the trees by dividing the frequency each feature was used at the 0^{th} level by the frequency each feature was a randomly sampled candidate. Features for the Dragon7 and ISIDA models were separately sorted descending on percent chosen at the first split point. The rank

ordering of the top forty Dragon7 features, the MLOG feature, and the top twenty ISIDA features between the β-lactamase and cruzain targets were comparatively analyzed.

RDKit features available at http://www.rdkit.org which are broadly related to weight, polarity, and aliphaticity were calculated for the entire unbalanced datasets. AMW, SlogP, and FractionCSP3 RDKit features were selected and each used to write single-feature models for the unbalanced β-lactamase dataset which classified compounds as putative aggregators if attribute values were greater than or equal to an optimized threshold tailored to that dataset established by iterative adjustment until the SE and SP of the resulting models were approximately equal. A model was built using the logic of all three optimized feature classifiers which predicted the class with at least two of the three classifiers agreeing. Finally, a random-forest model was built for the balanced β-lactamase dataset using each of these three features in addition to the highly prioritized ISIDA C-H feature. Modeling was carried out as previously discussed, with the exception that no AD was applied.

3. Results

3.1 Concentration Dependence

The frequency of active datapoints within variable AC_{50} concentration ranges was calculated for the raw detergent-free HTS assays to estimate the density of potential CACs (**Table 1**).

AC_{50} Concentration Bin	Target	# Datapoints	% Active Data
[10 μM, 0.1 mM)	β-lactamase	45,625	70.30
	cruzain	25,168	89.18
[1 μM, 10 μM)	β-lactamase	15,223	23.46
	cruzain	2,768	9.81
[0.1 μM, 1 μM)	β-lactamase	2,718	4.19
	cruzain	205	0.73
[10 nM, 0.1 μM)	β-lactamase	656	1.01
	cruzain	68	0.24
[1 nM, 10 nM)	β-lactamase	603	0.93
	cruzain	12	0.04
[0.1 nM, 1 nM)	β-lactamase	46	0.07
	cruzain	0	0.00

Table 1 - Aggregator concentration dependence for PubChem datasets
Brackets indicate inclusion of the given concentration, while parentheses indicate exclusion.

3.2 Dataset Concordance

Intra- and inter-dataset concordance was evaluated on duplicate entries which have the same compound tested in multiple assays to establish confidence in assay results and verify compatibility between HTS campaigns against the same target (**Table 2** and **Table 3**).

Target	HTS AIDs	# Duplicate Compounds	# Duplicate Sets	% Agreement	# Non-Agg Sets Agree	# Agg Sets Agree
β-lactamase	485341/485294	2,725	1,298	92.84	1,075	130
	585/584	1,496	648	91.05	519	71
cruzain	1476/1478	7,728	3,062	90.40	2,538	230

Table 2 - Intra-dataset concordance analysis
Duplicate sets are groups of instances in which a single compound has been tested in multiple assays. % Agreement indicates the percent of duplicate sets which unanimously agree in binary classification of the duplicate compound.

		585/584 Classification	
		Putative Aggregator	Non-Aggregator
485341/485294 Classification	Putative Aggregator	2,116	696
	Non-Aggregator	6,246	22,998

Table 3 - β-lactamase inter-dataset concordance analysis
Confusion matrix illustrating the classification agreement on compounds overlapping between both β-lactamase HTS campaigns.

3.3 Context Dependence

Chemical overlap and classification agreement between the β-lactamase and cruzain targets was evaluated to estimate the degree to which aggregation is a context dependent phenomenon (**Figure 4** and **Table 4**).

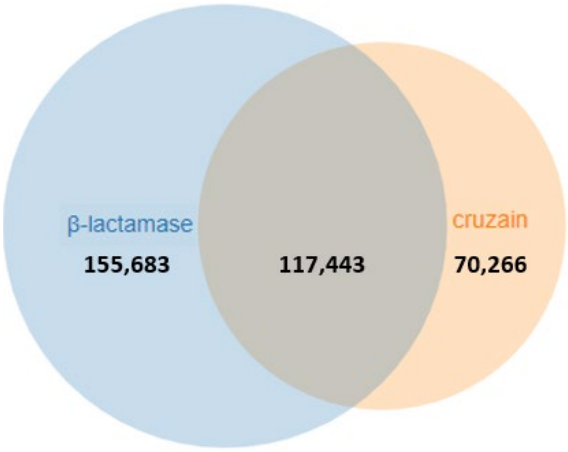


Figure 4 - Overlapping compounds tested between targets
Venn Diagram illustrating the number of curated compounds for the entire β-lactamase and cruzain datasets and the number of compounds tested against both targets.

		β-lactamase Classification	
		Putative Aggregator	Non-Aggregator
Cruzain Classification	Putative Aggregator	4,378	9,898
	Non-Aggregator	8,653	94,514

Table 4 - Inter-target concordance analysis
Confusion matrix illustrating the classification agreement for compounds overlapping between both curated target datasets.

3.4 Aggregator Advisor Analyses

3.4.1 Classification Agreement

The orthogonal sources used to collect compounds for the Aggregator Advisor[25] Rogues' Gallery were evaluated to determine whether classification inconsistencies exist for detergent-sensitive aggregators (**Table 5**).

Source ID	Reference	# Unique Compounds
1	Seidler J, McGovern SL, Doman TN, Shoichet BK, J Med Chem, 46, 4477-86 (2003)	47
2	Babaoglu et al, J Med Chem, 51, 2502-2011 (2008)	1,189
3	Shoichet Lab In House Data - Kristin Ziebart	9
4	Doak AK, Wille H, Pruisner SB and Shoichet BK, J Med Chem, 53, 4259-4265 (2010)	20
5	Coan KED and Shoichet BK, J.Am.Chem.Soc, 130, 9606-9612 (2008)	5
6	McGovern SL and Shoichet BK, J Med Chem 46, 1478-1483 (2003)	8
7	McGovern SL, Caselli E, Grigorieff N and Shoichet BK, J Med Chem, 45, 1712-1722 (2002)	24
8	Coan KE and Shoichet BK, Mol. BioSyst, 3, 208-213, (2007)	2
9	McGovern SL, Helfand, BT, Feng, B and Shoichet BK, J Med Chem, 46, 4265-4272 (2003)	5
10	Reddie KG, Roberts DR and Dore TM, J Med Chem, 49, 4857-4860 (2006)	1
11	Frenkel YV, Clark AD Jr, Das K, Wang UH, Lewi P, Jannssen PAJ and Arnold E, J Med Chem, 48, 1974-1983 (2005)	14
12	Giannetti AM, Koch BD, Browner MF, J Med Chem, 51, 574-480 (2008)	4
13	Liu HY, Wang Z, Regni C, Zou X, Tipton PA, Biochemistry, 43, 8662-8669 (2004)	1
14	Ferreira RS, Bryant C, Ang KKH, McKerrow JH, Shoichet BK, Renslo AR, J Med Chem, 52, 5005-5008 (2009)	11
15	Shoichet Lab In House Data - Allison Doak	23
16	Jadhav A et al, J Med Chem, 53, 37-51 (2010)	10
17	Mysinger M et al, PNAS, 109, 5517-5522 (2012)	1
18	Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, Keiser MJ, McKerrow JH, Maloney DJ, Irwin JJ and Shoichet BK, J Med Chem, 2010, 53, 4891-905.	11,232
?	?	2

Table 5 - Aggregator sources within the Advisor

The number of unique SMILES obtained from each orthogonal source used to build the Rogues' Gallery. No reference was provided for two SMILES.

Chemical overlap and classification agreement between the curated PubChem datasets and the curated Aggregator Advisor[25] Rogues' Gallery were evaluated to estimate the degree to which our definition of aggregation aligned with the Advisor's (**Figure 5** and **Figure 6**).

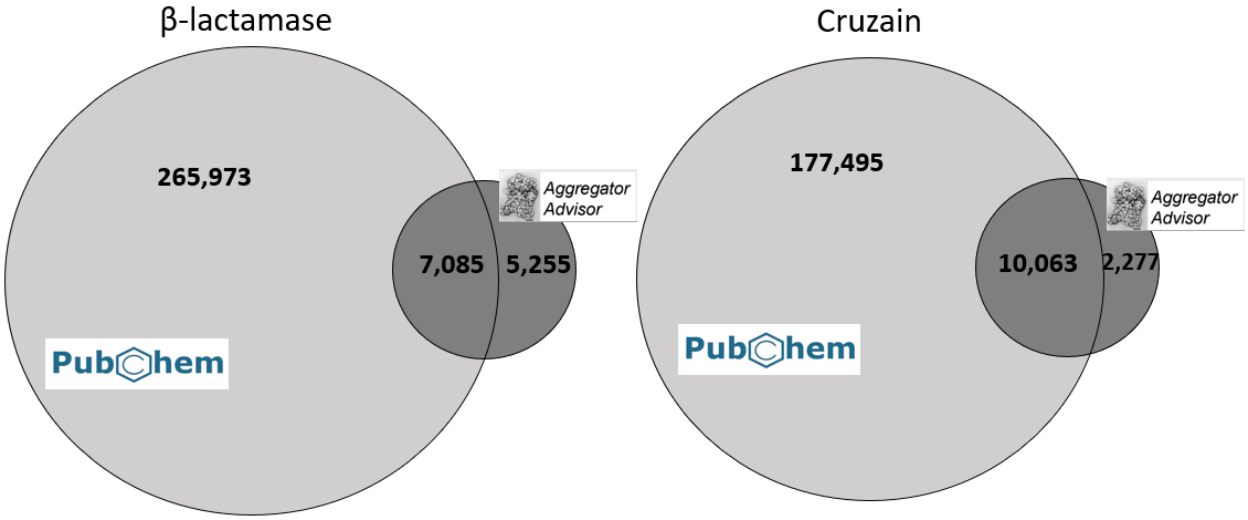


Figure 5 - Overlapping compounds used in the Advisor and tested in either target
Venn Diagrams illustrating the number of curated compounds for either target dataset from PubChem overlapping with curated compounds from the Aggregator Advisor. Not drawn to scale.

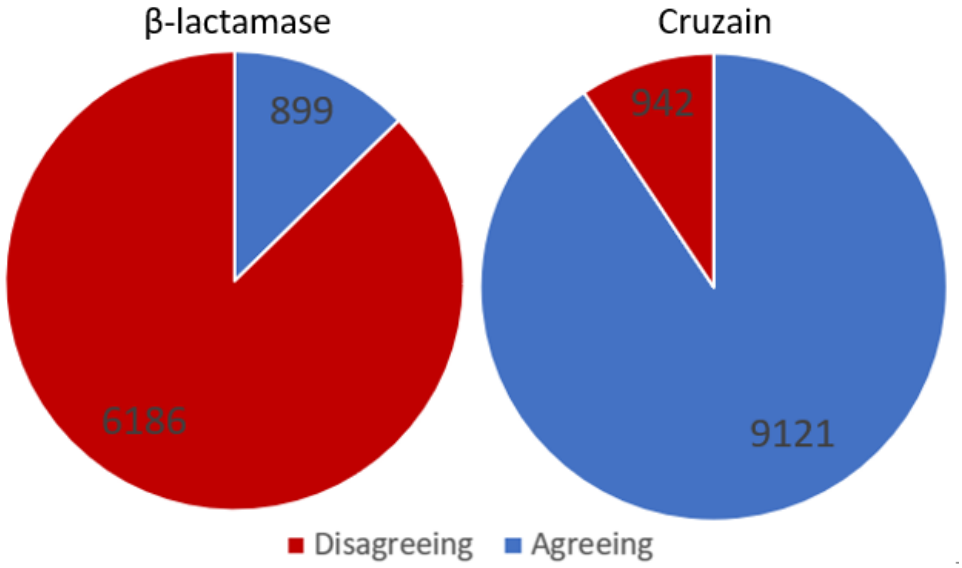


Figure 6 – Overlapping compound class agreement between target sets and the Advisor
Pie charts illustrating the percent classification agreement between compounds flagged as aggregators in the Aggregator Advisor and compounds from the β-lactamase and cruzain sets.

3.4.2 Programmatic Evaluation

The Aggregator Advisor[25] was evaluated to determine whether programmatic issues exist. A simple experiment was performed to evaluate the accuracy of the tool on its own reference database, the Rogues' Gallery. The logic used in the Aggregator Advisor to warrant an aggregation warning was applied to the SMILES in the Rogues' Gallery dataset after screening through the tool (**Table 6**).

# Unique SMILES	miLogP > 3?	TC > 0.8?
9,259	Yes	Yes
2,359	No	Yes
722	Yes	No
266	No	No

Table 6 - Rogues' Gallery Aggregators categorized by Advisor logic
The number of compounds from the Rogues' Gallery which obey or break either of the two criteria to be considered an aggregator by the Advisor tool. 808 Rogues found no matches to compounds within the Advisor. Approximately 17% of the SMILES from the largest source for the Rogues' Gallery did not find exact matches (TC = 1.0) through the Advisor tool.

Supplemental Material 1 showcases several examples of the Aggregator Advisor online tool either successfully identifying or failing to identify SMILES from its own Rogues' Gallery. This includes "CCCCSc3c(C)c2cc1OCOc1cc2nc3O", a SMILES with both miLogP > 3.0 and TC > 0.8, "CC(=O)C2=C(C)Nc1ncnn1C2c4cccc(OCc3ccccc3)c4" and "O=C3CC(c1ccco1)CC(Nc2ccccc2)=C3C(=O)Nc4ccccc4", two SMILES with miLogP > 3.0 and TC < 0.8, and "COC(\O)=c2/sc1cc(C)ccc1nc2C", a SMILES within the Rogues' Gallery which, in their words "*... has not been previously reported as an aggregator, or to be similar to an aggregator. (LogP 2.7)*". This response is generated for 222 unique SMILES from the Rogues' Gallery.

This last contradictory example SMILES, or CID 713501, was further examined on PubChem and was found to be both active and inactive in the same detergent-free assay against AmpC β-lactamase within two different HTS campaigns (**Figure 7**).

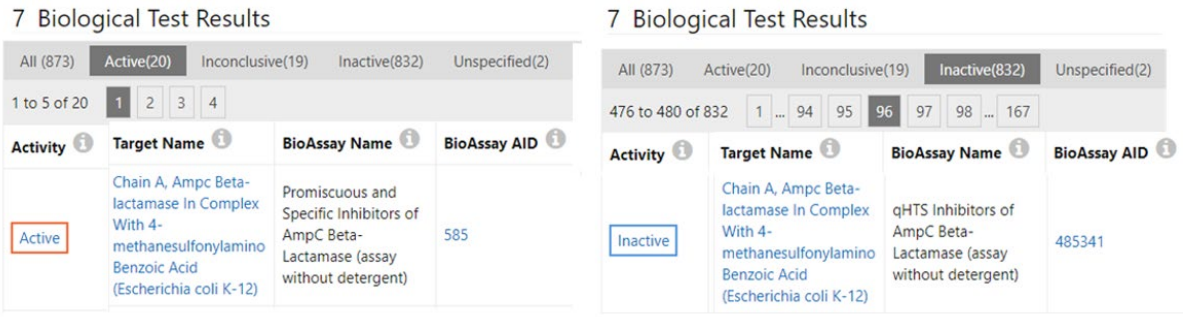


Figure 7 - Inconsistent results for an Aggregator tested in two β-lactamase assays
Cropped images from https://pubchem.ncbi.nlm.nih.gov/compound/713501#section=BioAssay-Results. Note that the maximum tested concentration for AID 585 (left) is nearly half that of AID 485341 (right).

3.5 Hit Dexter Analysis

The curated and classified chemical datasets for β-lactamase and cruzain were used as screening sets to interrogate the ability of the Hit Dexter tool[29] to distinguish the putative aggregator and non-aggregator classes (**Figure 8**).

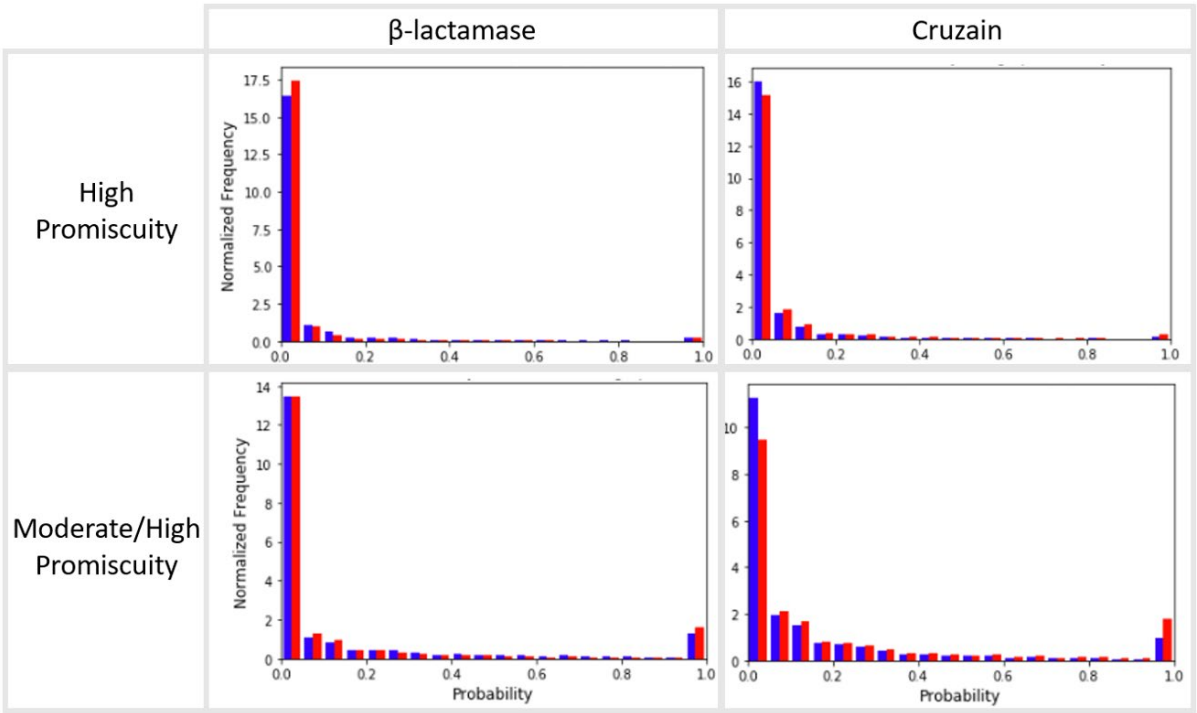


Figure 8 - Per-class Hit Dexter promiscuity probability distributions
High and moderate or high Hit Dexter promiscuity predictions were obtained from the entire, unbalanced and curated β-lactamase and cruzain datasets and were used to obtain histograms illustrating the difference in distributions between the non-aggregator (blue) and putative aggregator (red) classes. Histograms are normalized to possess a total area of 1. 20 bins used.

Hit Dexter generated two types of predictive warnings: "*Element types other than those present in the training data were detected. A result was generated but is probably unreliable.*" and "*The molecular weight is not between 250 and 900 Da. The prediction result may be unreliable.*" (**Table 7**).

Target	# Errors	# Non-Agg Warnings	# Agg Warnings	# Element Type Warnings	# MW Warnings
β-lactamase	21,745	21,264	476	65	21,680
cruzain	14,648	14,371	274	50	14,598

Table 7 – Hit Dexter error/warning frequency

3.6 MODI and QSPR Model Performance

The MODI was calculated to gauge whether balanced training sets maintained a tolerable frequency of activity cliffs[45,46], or compounds which are most similar to a compound of the opposing class at which a small change in structure results in a large change in activity (**Table 8**).

	Dragon7	ISIDA
β-lactamase	0.636	0.643
cruzain	0.632	0.645

Table 8 - Training set MODI calculations

MODI calculated for balanced, feature processed, non-normalized training sets.

Detergent-sensitive aggregation models were built against the β-lactamase and cruzain targets and evaluated for predictive accuracy (**Table 9** and **Table 10**, respectively).

Feature Set(s)	Type	SE	SP	Balanced Accuracy	NPV	PPV	Coverage
Dragon7	5FCV Model	0.752	0.749	0.75	0.751	0.75	1.0
	5FCV y-Rand	0.503	0.503	0.503	0.503	0.503	1.0
	Withdrawn Non-Aggregators	-	0.764	-	-	-	1.0
	Detergent-Resistant Actives	-	0.765	-	-	-	1.0
ISIDA	5FCV Model	0.702	0.756	0.729	0.717	0.742	1.0
	5FCV y-Rand	0.567	0.437	0.502	0.502	0.502	1.0

	Withdrawn Non-Aggregators	-	0.768	-	-	-	1.0
	Detergent-Resistant Actives	-	0.824	-	-	-	1.0
Consensus: Dragon7 + ISIDA	5FCV Model	0.773	0.802	0.787	0.78	0.794	0.834
	5FCV y-Rand	0.565	0.438	0.502	0.502	0.502	0.506
	Withdrawn Non-Aggregators	-	0.821	-	-	-	0.829
	Detergent-Resistant Actives	-	0.845	-	-	-	0.853

Table 9 - Predictive statistics of β-lactamase models

Feature Set(s)	Type	SE	SP	Balanced Accuracy	NPV	PPV	Coverage
Dragon7	5FCV Model	0.741	0.694	0.718	0.728	0.708	1.0
	5FCV y-Rand	0.505	0.491	0.498	0.498	0.498	1.0
	Withdrawn Non-Aggregators	-	0.686	-	-	-	1.0
	Detergent-Resistant Actives	-	0.589	-	-	-	1.0
ISIDA	5FCV Model	0.609	0.764	0.687	0.662	0.721	1.0
	5FCV y-Rand	0.158	0.845	0.502	0.501	0.505	1.0
	Withdrawn Non-Aggregators	-	0.741	-	-	-	1.0
	Detergent-Resistant Actives	-	0.642	-	-	-	1.0
Consensus: Dragon7 + ISIDA	5FCV Model	0.738	0.793	0.766	0.763	0.771	0.758
	5FCV y-Rand	0.16	0.841	0.50	0.497	0.505	0.495
	Withdrawn Non-Aggregators	-	0.786	-	-	-	0.749
	Detergent-Resistant Actives	-	0.684	-	-	-	0.629

Table 10 - Predictive statistics of cruzain models

3.7 Interpretation of QSPR Models

Features were sorted on frequency chosen at the first split point in the decision trees to approximately prioritize contribution to the prediction of aggregation. The top forty Dragon7 features were identified for both targets (**Table 11**) and the class-specific distributions for select features were visualized (**Figure 9**).

Dragon7 Descriptor	Feature Description	% Chosen: β-lactamase	% Chosen: cruzain
CATS2D_08_AP	CATS2D Acceptor-Positive at lag 08	37.2	22.9
F03[O-P]	Frequency of O - P at topological distance 3	32.4	55.7
CATS2D_04_DP	CATS2D Donor-Positive at lag 04	31.9	41.1
F10[F-F]	Frequency of F - F at topological distance 10	24.2	27.1
T(P..P)	sum of topological distances between P..P	23.1	40.5
AAC	mean information index on atomic composition	100.0	14.3
AMW	average molecular weight	91.8	0.0
F03[C-C]	Frequency of C - C at topological distance 3	90.7	7.2
SpMin5_Bh(m)	smallest eigenvalue n. 5 of Burden matrix weighted by mass	85.3	13.5
SpMin4_Bh(m)	smallest eigenvalue n. 4 of Burden matrix weighted by mass	79.9	NA
ATS7p	Broto-Moreau autocorrelation of lag 7 (log function) weighted by polarizability	73.8	NA
SpMin8_Bh(s)	smallest eigenvalue n. 8 of Burden matrix weighted by I-state	68.8	1.7
F06[C-C]	Frequency of C - C at topological distance 6	66.4	1.2
SpMin6_Bh(e)	smallest eigenvalue n. 6 of Burden matrix weighted by Sanderson electronegativity	61.8	NA
CIC1	Complementary Information Content index (neighborhood symmetry of 1-order)	60.8	1.1
CATS2D_08_DP	CATS2D Donor-Positive at lag 08	59.5	0.0
ATS2e	Broto-Moreau autocorrelation of lag 2 (log function) weighted by Sanderson electronegativity	57.3	NA
F08[F-Cl]	Frequency of F - Cl at topological distance 8	55.3	21.8
P_VSA_MR_1	P_VSA-like on Molar Refractivity, bin 1	53.3	1.8
N-075	R--N--R / R--N--X	53.0	0.0
F10[O-Br]	Frequency of O - Br at topological distance 10	51.0	0.0
H%	percentage of H atoms	47.7	0.0
F09[O-Cl]	Frequency of O - Cl at topological distance 9	44.2	9.7
ATSC6p	Centred Broto-Moreau autocorrelation of lag 6 weighted by polarizability	44.2	0.0
CATS2D_08_DN	CATS2D Donor-Negative at lag 08	41.0	1.5
ATSC8m	Centred Broto-Moreau autocorrelation of lag 8 weighted by mass	38.4	NA
SM09_AEA(dm)	spectral moment of order 9 from augmented edge adjacency mat. weighted by dipole moment	38.3	NA
TPC	total path count	34.0	NA
CATS2D_06_PL	CATS2D Positive-Lipophilic at lag 06	33.3	0.0
T(Cl..I)	sum of topological distances between Cl..I	31.5	NA
AVS_Dt	average vertex sum from detour matrix	29.4	NA
ATSC2m	Centred Broto-Moreau autocorrelation of lag 2 weighted by mass	28.8	0.0
F09[C-C]	Frequency of C - C at topological distance 9	27.0	4.3

HyWi_H2	hyper-Wiener-like index (log function) from reciprocal squared distance matrix	26.3	2.9
H-047	H attached to C1(sp3)/C0(sp2)	25.9	0.0
Wap	all-path Wiener index	24.5	5.9
C-016	"=CHR"	23.4	NA
F07[F-F]	Frequency of F - F at topological distance 7	23.2	11.0
CATS2D_05_AP	CATS2D Acceptor-Positive at lag 05	22.2	12.5
P_VSA_MR_4	P_VSA-like on Molar Refractivity, bin 4	21.9	3.6
CATS2D_04_DD	CATS2D Donor-Donor at lag 04	0.8	90.6
Eig12_EA(dm)	eigenvalue n. 12 from edge adjacency mat. weighted by dipole moment	NA	85.9
Eig15_EA(dm)	eigenvalue n. 15 from edge adjacency mat. weighted by dipole moment	10.9	83.6
F02[O-O]	Frequency of O - O at topological distance 2	0.0	80.5
F07[F-Cl]	Frequency of F - Cl at topological distance 7	NA	76.5
Eig13_EA(dm)	eigenvalue n. 13 from edge adjacency mat. weighted by dipole moment	17.4	74.3
T(N..P)	sum of topological distances between N..P	15.3	65.0
Eig14_EA(dm)	eigenvalue n. 14 from edge adjacency mat. weighted by dipole moment	9.6	58.8
CATS2D_02_AA	CATS2D Acceptor-Acceptor at lag 02	0.0	54.9
B09[C-C]	Presence/absence of C - C at topological distance 9	14.1	50.0
T(O..P)	sum of topological distances between O..P	15.7	48.3
B08[C-C]	Presence/absence of C - C at topological distance 8	14.2	42.2
F03[O-Cl]	Frequency of O - Cl at topological distance 3	6.3	41.5
T(S..P)	sum of topological distances between S..P	7.6	39.3
B10[C-C]	Presence/absence of C - C at topological distance 10	8.4	39.1
F06[O-P]	Frequency of O - P at topological distance 6	NA	38.9
CATS2D_08_DD	CATS2D Donor-Donor at lag 08	0.0	37.2
F07[O-P]	Frequency of O - P at topological distance 7	NA	35.3
CATS2D_07_DA	CATS2D Donor-Acceptor at lag 07	0.0	34.8
SdsssP	Sum of dsssP E-states	3.1	34.4
CATS2D_02_DL	CATS2D Donor-Lipophilic at lag 02	0.0	33.9
B07[C-C]	Presence/absence of C - C at topological distance 7	NA	33.5
F03[C-I]	Frequency of C - I at topological distance 3	8.5	30.3
F03[Cl-Cl]	Frequency of Cl - Cl at topological distance 3	11.4	29.0
SssNH	Sum of ssNH E-states	0.0	26.3
D/Dtr04	distance/detour ring index of order 4	5.1	26.1
F06[C-I]	Frequency of C - I at topological distance 6	NA	25.3
F05[O-O]	Frequency of O - O at topological distance 5	1.4	25.2
F03[N-F]	Frequency of N - F at topological distance 3	1.6	24.4
CATS2D_08_AN	CATS2D Acceptor-Negative at lag 08	0.0	24.2
F05[F-F]	Frequency of F - F at topological distance 5	0.0	23.3
F07[C-P]	Frequency of C - P at topological distance 7	1.4	23.3
SsSH	Sum of sSH E-states	6.1	22.9
CATS2D_05_DA	CATS2D Donor-Acceptor at lag 05	0.0	22.8
WiA_Dt	average Wiener-like index from detour matrix	9.9	22.8

Table 11 - Model-prioritized Dragon7 features

Grey features overlap in the top forty for both targets, blue features are ordered with descending importance for β-lactamase, and orange features are ordered with descending importance for cruzain. "NA" indicates that this feature was filtered during pre-processing.

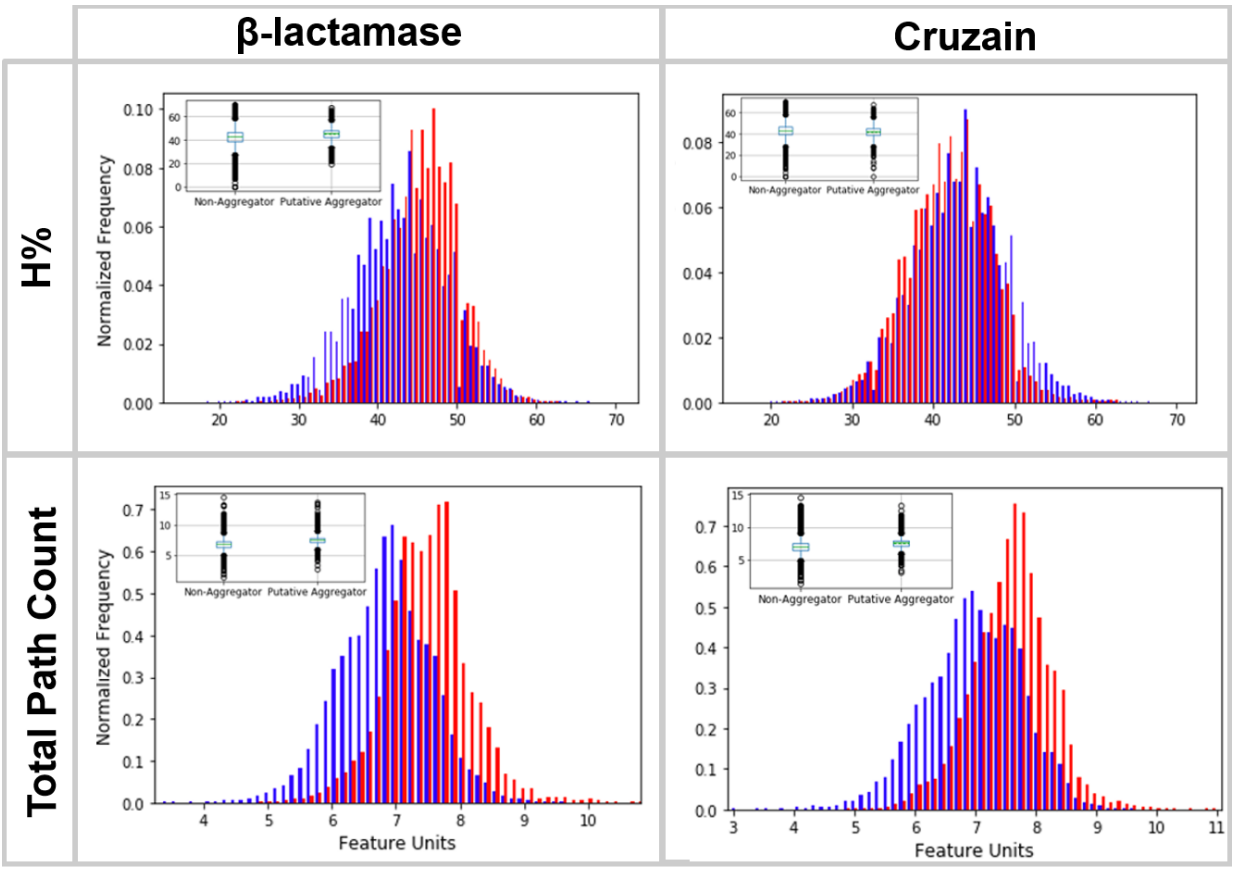


Figure 9 - Per-class distributions for select Dragon7 model-prioritized features
Histograms and inset box plots illustrating distributions for the non-aggregator (blue) and putative aggregator (red) classes on Dragon7 features selected based on interpretability which were calculated from the entire unbalanced, curated β-lactamase and cruzain datasets. Histograms are normalized to possess a total area of 1. In most cases 100 bins were used. Distributions for additional features are given in **Supplemental Material 8**.

The Aggregator Advisor[25] uses logP as a key metric to flag potential aggregators. To probe whether logP is predictive of aggregation and effectively distinguishes the non-aggregator and putative aggregator classes, the MLOGP Dragon7 feature was examined for model prioritization and class-specific distribution (**Figure 10**). Neither target model chose the MLOGP feature at the first split point in any tree. The β-lactamase model chose the feature 1.32% of the time at the second split level and 2.77% of the time at the third, while the cruzain model chose the feature 1.82% at the second split point and 1.38% of the time at the third.

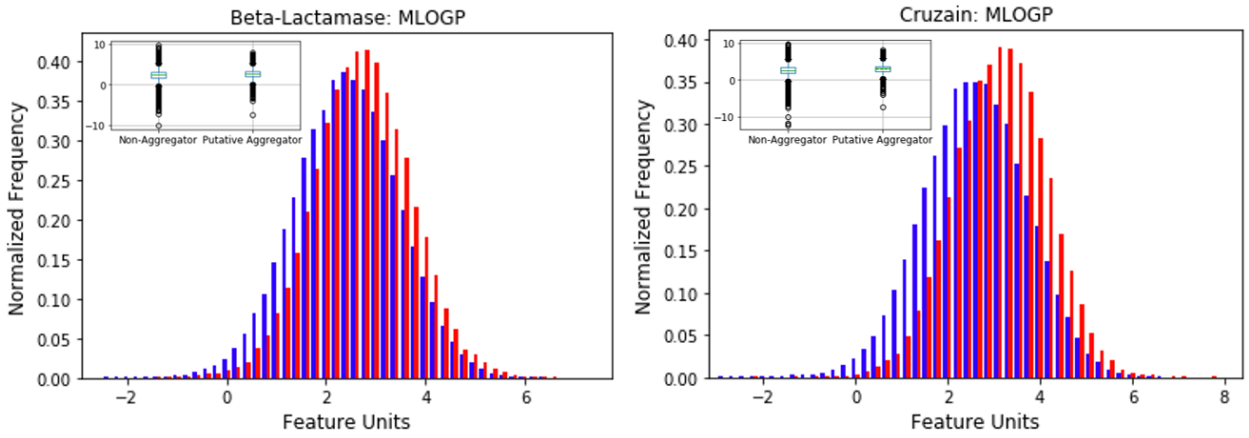


Figure 10 - Per-class distributions for MLOGP

Histograms and inset box plots illustrating distributions for the non-aggregator (blue) and putative aggregator (red) classes on the MLOGP Dragon7 feature calculated from the entire unbalanced, curated β-lactamase and cruzain datasets. Histograms are normalized to possess a total area of 1. 100 bins were used.

The top twenty ISIDA features were identified for both targets (**Table 12**) and the class-specific distributions for the C-H feature was visualized (**Figure 11**).

ISIDA Descriptor	% Chosen: β-lactamase	% Chosen: cruzain
H-C-C-S-C=S	63.9	41.1
H-C-C-N-C-N=O	52.2	56.2
H-C-C-C-N=C-H	50.4	36.1
F-C-C-N=C-C-H	37.4	32.1
F-C-C-N-C-N	36.8	51.1
C-N-C=C-C-N-H	82.4	NA
H-N-C-C-C-C-O	65.4	NA
N-C=C-C-N-S=O	62.7	NA
C-H	58.2	0.7
H-C-N-C=C-P-O	56.8	NA
F-C-C-N-S=O	53.7	30.7
H-N-C-C-N=N	46.9	NA
H-C-N-N-S=O	44.6	NA
H-C-S-C-C=C-S	41.3	NA
H-C-C-C=C-N=N	39.3	NA
H-C-N-N=N-C=N	38.0	1.4
C=C-S-C-C-C-N	33.0	0.0
C-C-C-N-N-C-H	32.6	1.9
C-C-C=N-C-C=O	32.4	NA
H-C=C-O-C-C-N	31.9	0.8
N-C-C-C=O	0.0	86.8
H-C-C-C-S-C=S	NA	73.3

N-C-C-C-C-O	0.0	64.4
N-C-C-C-O	6.3	59.4
O-P-O-P	NA	53.8
F-C-C-C-C-H	3.3	52.7
C-C-N-O-H	NA	42.4
C-C-C-C-O-C=N	NA	38.8
H-N=C-N	1.8	37.9
H-C-O-P-S-C-H	NA	36.5
C-C-C-C=N-O	0.0	35.8
C-C-C=C-C-C-F	15.8	34.7
H-N-C-C-C-O	15.3	34.1
C=C-C-C-N-C-O	NA	33.1
C-C-C-C=C-N-C	0.0	31.6

Table 12 - Model-prioritized ISIDA features

Grey features overlap in the top twenty for both targets, blue features are ordered with descending importance for β-lactamase, and orange features are ordered with descending importance for cruzain. "NA" indicates that this feature was filtered during pre-processing.

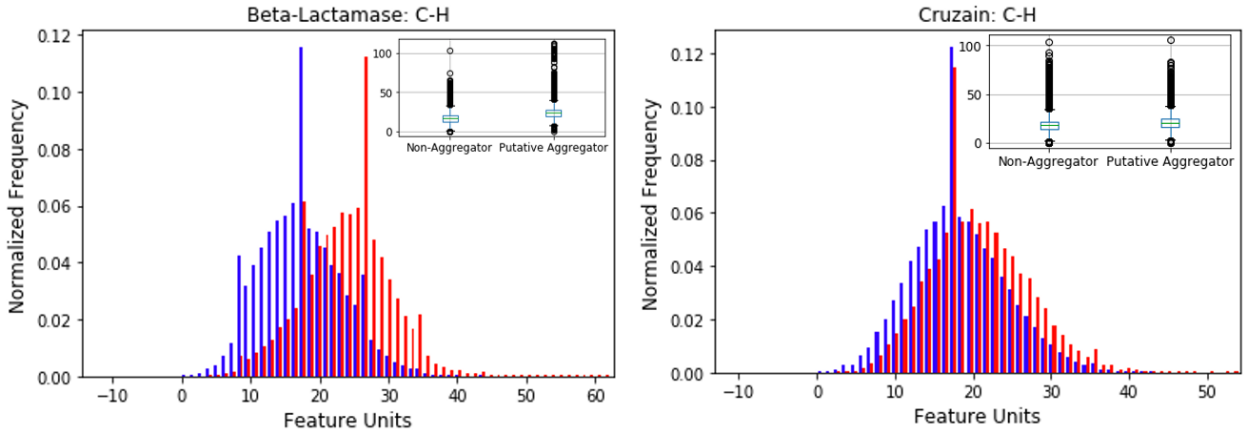
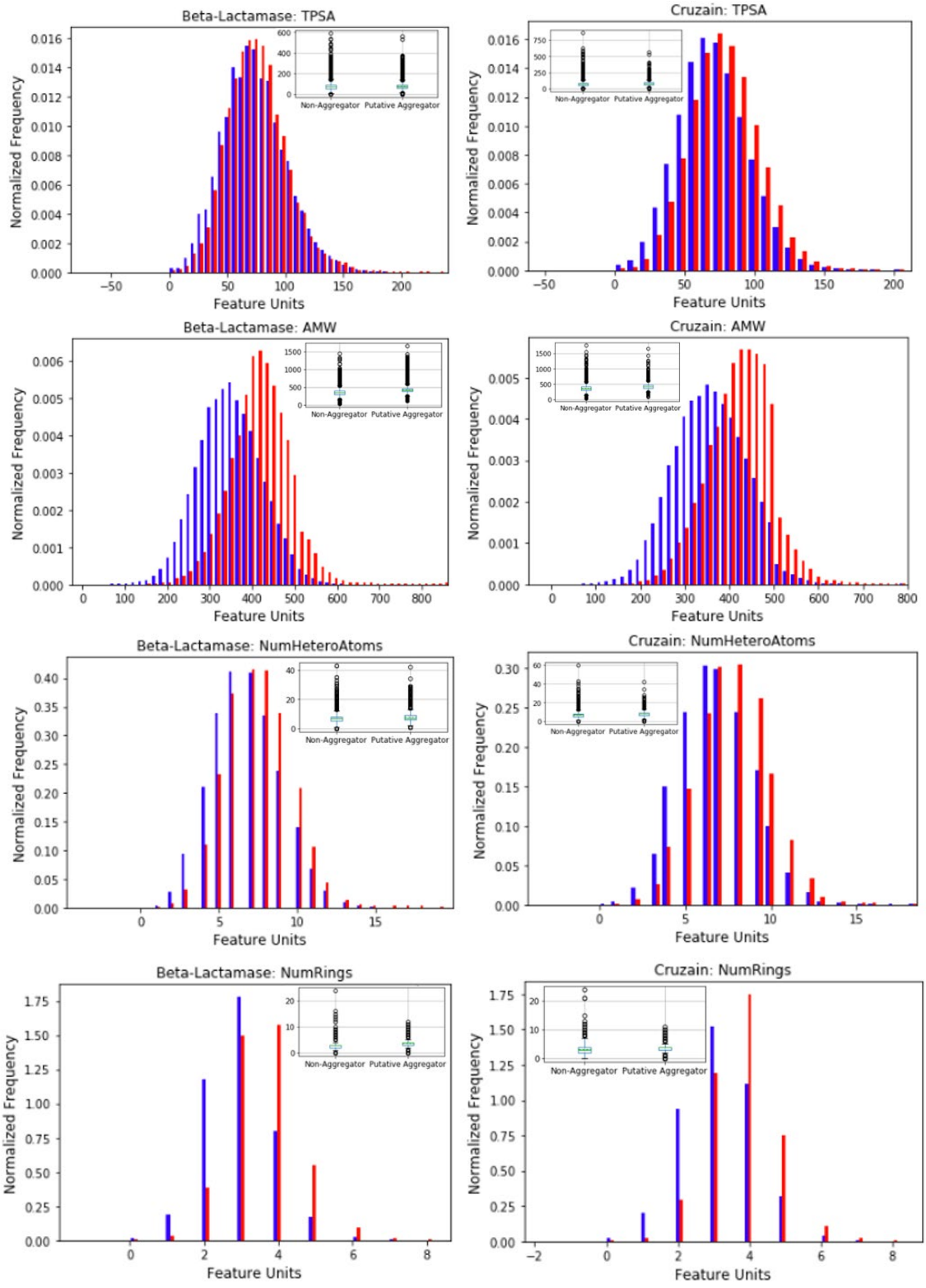


Figure 11 - Per-class distributions for C-H bond count

Histograms and inset box plots illustrating distributions for the non-aggregator (blue) and putative aggregator (red) classes on the ISIDA C-H feature which was calculated from the entire unbalanced, curated β-lactamase and cruzain datasets. Histograms are normalized to possess a total area of 1. 100 bins were used.

Insights gained through Dragon7 and ISIDA feature analysis were used to test whether aggregation models could be simplified to rules for a small number of RDKit features broadly related to weight, polarity, and aliphaticity. The class-specific distributions for select RDKit features were visualized (**Figure 12**).


Beta-Lactamase: TPSA
Cruzain: TPSA
Beta-Lactamase: AMW
Cruzain: AMW
Beta-Lactamase: NumHeteroAtoms
Cruzain: NumHeteroAtoms
Beta-Lactamase: NumRings
Cruzain: NumRings
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator

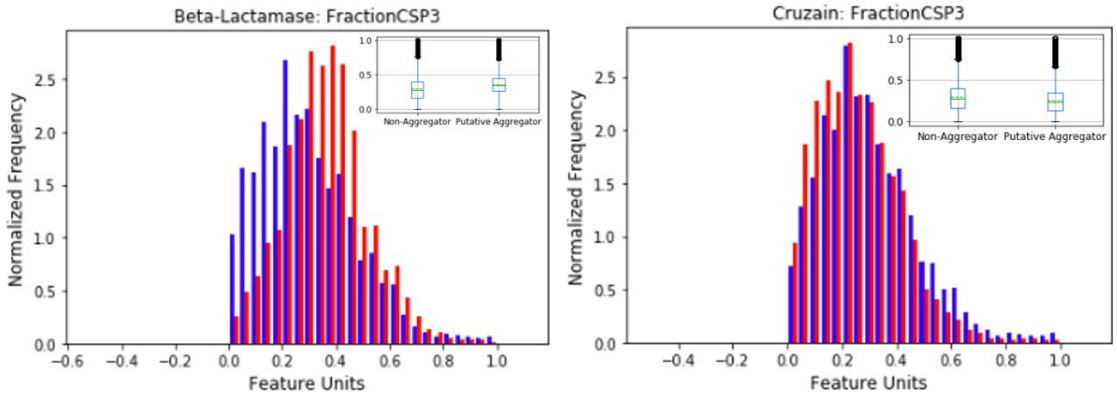


Figure 12 - Per-class distributions for select RDKit features
Histograms and inset box plots illustrating distributions for the non-aggregator (blue) and putative aggregator (red) classes on select RDKit features calculated from the entire unbalanced, curated β-lactamase (left column) and cruzain (right column) datasets. Histograms are normalized to possess a total area of 1. In most cases 100 bins were used.

To test whether AMW, SlogP, and FractionCSP3 features adequately approximate detergent-sensitive aggregation, each feature was used to build single-feature models which classified compounds based on whether attributes were greater than or equal to an optimized threshold. Additionally, a model was built using the logic of all three optimized classifiers which predicted the class with at least two of the three classifiers agreeing (**Table 13**).

RDKit Feature	Optimized Threshold	SE	SP	Balanced Accuracy	NPV	PPV
AMW	385	0.694	0.69	0.692	0.948	0.217
FractionCSP3	0.3175	0.607	0.606	0.606	0.926	0.16
SlogP	3.175	0.594	0.596	0.595	0.922	0.154
All three	All optimized	0.709	0.692	0.70	0.95	0.222

Table 13 - Predictive statistics for RDKit β-lactamase models with optimized thresholds
Thresholds were optimized and classifiers were built specifically for the entire unbalanced β-lactamase dataset.

Finally, to compare the consensus models built from the complete Dragon7 and ISIDA sets with models built with a refined selection of features, a random-forest classifier was built using AMW, SlogP, and FractionCSP3 RDKit features and the highly prioritized ISIDA C-H feature (**Table 14**).

Type	SE	SP	Balanced Accuracy	NPV	PPV
5FCV Model	0.682	0.689	0.686	0.685	0.687
5FCV y-Rand	0.476	0.522	0.499	0.499	0.499
Withdrawn Balancing Non-Aggregators	-	0.72	-	-	-

Table 14 - Predictive statistics for the refined, hybrid-feature β-lactamase RF model
Models were built specifically for the balanced β-lactamase dataset. No AD was applied.

4. Discussion and Conclusions

Our collection of dose-response HTS data reporting on detergent-sensitive aggregation is the largest used for cheminformatics analysis to-date. This makes us uniquely positioned to develop novel models, as the strength and prospective accuracy of statistical predictors is often a function of the wealth of data used in training.[25] Considering that the majority of aggregators have CACs in the low micromolar range[12,19], these screens performed on concentration gradients ranging between 1.8 nM and 57.10 μM give us high sensitivity to detect artifact-like activity and allow our training data to be applicable to the majority of relevant doses tested in standard HTS.

Our estimation of the density of aggregator CACs demonstrated that there are insufficient data to generate a series of models which predict aggregators at variable HTS concentration ranges. Only the two largest tested concentration ranges, [10 μM, 0.1 mM) and [1 μM, 10 μM), contained a significant enough portion of the active data to adequately model concentration-specific aggregation, with each of the other ranges containing less than 5% of the active data for both target datasets. Therefore, models presented in this work are especially enriched in aggregators with CACs between 1 μM and 100 μM, and these models will be especially suited to predict future aggregators with CACs in this range.

There is no field consensus in quantitatively defining aggregation based on detergent-sensitive assay interpretation in lieu of physical characterization. For example, Shoichet *et al.*[24] labeled aggregators as detergent-sensitive compounds which inhibit three distinct enzymes in a

time-dependent manner, while NCGC[12] claims that aggregators should be flagged if there is a greater than 20% reduction in activity as a result of adding detergent at a given concentration and that "*rightward shifts in IC_{50} > 3x should be considered likely aggregate-dependent effects*".[12] Using the proxy of biochemical sensitivity to detergent in the absence of physical evidence of colloidal particles is imperfect, and no definition, including our own, is without exceptions. This experimental uncertainty is inherent in any QSAR/QSPR modeling endeavor. Therefore, we refer to our aggregators as "putative" to establish the qualification that this classification is extrapolated through biochemical assay and may, in some cases, be incorrect. For example, our definition misclassifies colloidal aggregates which have no inhibitory effect on, or enhance, the activity of a specific target protein[3], are not perturbed by detergent, or have low affinity to adsorb to β-lactamase or cruzain proteins.[3,10,18] Technically, compounds which are detergent-insensitive may either be inert with and without detergent or be active with and without detergent. However, while defining the non-aggregator class, we observed that only tens of compounds were detergent-resistant actives, while hundreds of thousands of compounds were completely inert. To prevent confusion of the modeling endpoint with compounds which have drastically variable experimental activities, we chose to use the large set of inert compounds to define the non-aggregator class and maintained the detergent-resistant actives for external testing.

We demonstrate the pervasiveness of putative aggregators in HTS, as more than 11% of the curated β-lactamase and cruzain datasets showed some degree of detergent-sensitive activity. This surpasses prior speculations that aggregators may plague up to 5% of some screening libraries[12] and emphasizes the hazard of aggregators in HTS.

Concordance analysis established strong internal confidence for the HTS campaigns, as each had more than 90% intra-dataset agreement on duplicate entries. However, inter-dataset

agreement between the two HTS campaigns against β-lactamase was less confirmatory. Only about 78% of compounds overlapping between the sets were assigned the same classification, with approximately 90% of disagreeing compounds being labeled as non-aggregators in HTS AIDs 485341/485294 and putative aggregators in HTS AIDs 585/584. This runs contrary to expectation, as the former HTS was performed at a maximal concentration approximately twice that of the latter. Regardless, we deemed this degree of concordance suitable considering that these assays were performed under identical conditions and proceeded to merge datasets to ultimately form a single β-lactamase aggregation model.

Our analyses reinforce prior notions that aggregation is a heavily context-dependent phenomenon.[12] More than 22,000 unique NCGC compounds were screened against both the β-lactamase and cruzain targets and were labeled aggregators in at least one of the two datasets. However, more than 80% of these aggregators were non-aggregators under the second set of assay conditions, which varied significantly by buffer system and pH, protein concentration, and detection technology. Considering this degree of discordance, and that aggregation is highly context dependent, we maintained the two sets separately to model in parallel with refined assay contexts.

We highlight major flaws in the classification definition and programmatic functionality of the Aggregator Advisor tool[25]. First, we observed that the historical data collected for the Advisor were obtained from 18 orthogonal sources, some of which originate from unpublished reserves. Based on the verbiage of the original publication[25], there is no consistent quantitative threshold which was used to classify compounds as aggregators from one data source to another. Instead, aggregators were classified if they broadly "*caused inhibition that was reversible by detergent*". For example, one source[11] classified compounds as aggregators if they "*had a maximum response*

greater than 40% inhibition" against cruzain sensitive to the addition of detergent, while another source[21] considered compounds aggregators "*if they produced a curve class of -3 or better … over a concentration range from 4 nM to 30 μM*" against AmpC β-lactamase which was sensitive to the addition of detergent. As aggregation has been perceived as a context-dependent phenomenon[12,19], these classification inconsistencies make it difficult to discern whether a compound in the Rogues' Gallery aggregates in a certain assay and causes inhibition of a certain target. Second, the public data used in this study corresponds to the two sources which contribute more than 98% of the compounds in the Rogues' Gallery. However, the Gallery only contains 1,189 aggregators from β-lactamase HTS AIDs 585/584 and 11,232 aggregators from cruzain HTS AIDs 1476/1478, which from our interpretations contained 13,583 and 24,660 curated problematic detergent-sensitive compounds, respectfully. Additionally, the β-lactamase HTS AIDs 485341/485294 resulted in 25,586 unique curated putative aggregators, but these data were never equipped to support the Rogues' Gallery. This resulted in one example compound within the Rogues' Gallery, CID 713501, which is labeled in PubChem to have contradictory experimental evidence being both active and inactive in separate PubChem AmpC β-lactamase screens without detergent. This example may be representative of many low confidence or misclassified aggregators within the Rogues' Gallery.

While comparing our classified datasets to the compounds in the Advisor we found that the majority of the Rogues' Gallery existed as a minor subset of our curated sets. However, there was a stark difference in the classification of compounds which overlap between the Advisor and the β-lactamase set. More than 87% of these compounds were considered aggregators by the Advisor but were inert in all PubChem HTS data against β-lactamase used in this study. This was likely a result of the Adviser not considering information from the HTS AIDs 485341/485294. On the

contrary, there was more than 90% agreement in the classification of putative aggregators between our cruzain set and the overlapping compounds in the Rogues' Gallery. This is likely because a significant majority of the compounds in the Gallery derive from the cruzain HTS campaign.[47] Therefore, it would seem that the Advisor is more well positioned to perform advisement of compounds which aggregate in conditions tested specifically in the cruzain HTS AIDs 1476/1478, but may not correctly flag compounds from other assay conditions or against different targets such as β-lactamase.

While probing programmatic issues with the Aggregator Advisor we found that the provided python batch script does not allow the user to indicate a specified Affinity Range. This, in addition to our observations that the outcome of any given SMILES query is not dependent on whether the Affinity Range is set on *"< 0.1 μM*", "*0.1-10 μM*", or "*> 10 μM*", provides evidence that the tool does not include information on compound concentration to inform its analysis. Next, due to some programmatic error, the absence of certain data in the TC search algorithm, or the non-standardization of SMILES, an alarming 808 compounds within the Rogues' Gallery cannot identify themselves within the Aggregator Advisor tool and 222 SMILES from the Rogues' Gallery cause the Advisor to respond that the compound *"…has not been previously reported as an aggregator, or to be similar to an aggregator.*". About 26% of the SMILES from the Rogues' Gallery do not possess both criteria used to consider aggregators by the Advisor, and more than 20% of the SMILES do not have miLopP > 3.0, which may indicate that this criterion does not effectively identify aggregators. Altogether, these issues with the Aggregator Advisor necessitate a more reliable predictive model for colloidal aggregation.

The Hit Dexter tool (version 1.0)[29] failed to discriminate putative aggregators from non-aggregators based on the per-class distribution of promiscuous probabilities. We assume that the

frequency of promiscuous compounds in the aggregating class should be significantly higher, theoretically 100%, than the non-aggregating class, especially considering that aggregation is the most frequent cause of non-specific activity in HTS assays.[4,10] Therefore, the Hit Dexter tool should rank the promiscuity of the putative aggregator class well above that of the non-aggregator class for both datasets. Per-class Hit Dexter promiscuity probability distributions illustrate an indiscriminately small separation between the non-aggregator and putative aggregator classes, with approximately equal median and mean values for both the high and moderate or high promiscuity endpoints in both targets. A small exception to this is in the case of the moderate or high promiscuity value for the cruzain dataset, in which there is approximately a 5% increase in the mean promiscuity of the aggregator class over the non-aggregator class. Additionally, two errors occurred while screening. There were 50 or more observations per target in which predictions were not reliable due to having unrecognized element types, and there were more than 14,500 unreliable predictions per target due to compounds being outside of the training set range of 250 and 900 Da. These two errors represent significant flaws of the tool, as most HTS libraries contain compounds which fall outside of this weight range. We note that we did not apply an applicability domain to these predictions due to computational demand, which may have altered predictive results. During the writing of this work a second version of the Hit Dexter tool was released (2.0) which was trained on more extensive dose-response data on interference mechanisms.[30] We did not evaluate the accuracy of version 2.0 on our datasets but suspect that this version might also fail to prioritize the aggregating class as promiscuous as a result of combining training set data related to many types of assay interference, therefore diluting and confusing the endpoint of the model and lowering predictive accuracy towards this specific mechanism.

We chose Dragon7 and ISIDA features to describe our compounds because they represent orthogonal feature generation techniques. Dragon7 is the state of the art in the description of whole-molecule physicochemical features, and ISIDA exhaustively describes the frequency, not just existence, of molecular substructures. The use of exclusively 2D features prevented us from discriminating rotomers or stereoisomers. These techniques generated massive numbers of descriptors in order to gain maximal class distinction for modeling, including 4,500 Dragon7 features and a sparse matrix of more than 29,000 ISIDA features.

Feature processing was performed to lighten the computational demand and simplify feature selection during machine learning. Binary statistical classifiers such as random forest are often sensitive to highly imbalanced training sets which have a disproportionate number of compounds between classes.[48] To address this, we applied a balancing method to undersample the non-aggregator class which uses distance-based and stochastic components. To match the size of the putative aggregator class, the nearest 50% non-aggregators to the putative aggregator class, the 25% furthest non-aggregators, and 25% random remaining non-aggregators were selected. This undersampling protocol theoretically retains compounds close to the interface between the two classes to inform the model about edge cases where the distinction between classes is small, compounds where the distinction between classes is massive, and compounds which are randomly distributed that maximize the diversity of the non-aggregator class. Feature processing after balancing maintained a massive number of informative descriptors, more than 800 for Dragon7 and more than 1,000 for ISIDA. We used a stricter correlation cutoff for ISIDA to significantly reduce the size of the sparse matrix and lower the likelihood of fragment redundancies due to a fragment existing as a subset of another.

Dataset MODI values indicated that our modeling datasets contained high but tolerable frequencies of activity cliffs. Datasets with MODIs significantly less than 0.65 may not be able to produce predictive models due to a saturation of activity cliffs which increase the difficulty for the statistical technique to discern characteristics which differentiate compounds between classes.[40] Because our balancing algorithm maintained the 50% non-aggregators closest to the opposing class, we anticipated a notable number of activity cliffs in our datasets. Each of our modeling dataset MODI values approached 0.65, falling in the range between 0.632 and 0.645. Therefore, we pursued the modeling of these data.

Our QSPR models proved to be rigorous and extensible, as they demonstrated high predictivity of detergent-sensitive aggregation for AmpC β-lactamase and cruzain-specific protein targets, with 5FCV balanced accuracies of 0.787 and 0.766, respectively. We chose a random forest statistical modeling approach for its ease of interpretation and lightweight computational demand.[43,44] Y-randomized controls verified that no model overfit the training datasets, as the 5FCV balanced accuracies for all y-randomized models fell within the range of 0.498 and 0.503, approximating the accuracy of a random guess. Performing consensus predictions using models generated from whole-molecule physicochemical features and fragment features allowed us to increase predictive confidence and establish a simple method for AD, as evidenced by the approximate 4% to 8% increase in 5FCV balanced accuracies when using consensus models instead of the single Dragon7 and ISIDA feature sets alone. Indeed, SE, SP, balanced accuracy, NPV, and PPV were all increased in the consensus models over the single feature set models for both targets, with the exception that the cruzain Dragon7 model had a 0.3% higher 5FCV SE than the consensus model. However, increased predictivity came at the expense of reduced coverage due to AD implementation, which caused up to a 17% reduction in the coverage of the β-lactamase

model and between about a 25% and 37% reduction for the cruzain screening sets. It seemed that the Dragon7 features contributed more to the classification of putative aggregators, with the 5FCV SE of the Dragon7 models between 5% and 13.5% greater than that of the ISIDA models, while the ISIDA features had a larger contribution to the identification of non-aggregators, with 5FCV SP between 0.5% and 7% greater than that of the Dragon7 models. Overall, the Dragon7 models outperformed the ISIDA models with 5FCV balanced accuracies between 2% and 3% higher, which may be a result of the sparsity of the ISIDA feature space. Additionally, the consensus β-lactamase model outperformed the cruzain model with a 5FCV balanced accuracy about 2% higher and coverage more than 7% greater for each screening set, likely a result of using a larger β-lactamase training set composed of high-confidence duplicate classifications between two separate HTS campaigns. Consensus models showed high SP in validation screens of hundreds of thousands of external non-aggregator compounds, with the β-lactamase and cruzain models correctly classifying non-aggregators about 82% and 74% of the time, respectively. Additionally, these models demonstrated strong coverage of external non-aggregator datasets indicating a wealth of model chemical diversity, as the β-lactamase and cruzain screening sets were about 83% and 75% covered, respectively. Finally, about 85% of the detergent-resistant actives fell within the AD of the β-lactamase consensus model, which correctly classified more than 84% as being non-aggregators. However, the cruzain consensus model struggled to recognize the small set of detergent-resistant actives, with less than 70% correctly predicted out of the 63% within the AD. This may indicate that the cruzain model has insufficient information on specific, true-positive compounds in HTS. However, predictions performed in conjunction with both models may have higher likelihood of being correctly classified in more generalized assay contexts.

QSAR/QSPR model interpretation is often difficult or impossible due to the complicated nature of machine learning algorithms or the expanse and diversity of training datasets. Even the logical algorithm of random forest becomes obscure when building five iterations of 1,000-tree forests. However, we managed to probe the complicated SIR of these models to identify features correlated to aggregation in the context of the β-lactamase and cruzain screens by studying the frequency that features are the most predictive single-model classifiers in tree samplings. Some of the top forty Dragon7 features for the β-lactamase model related to average molecular weight, which was chosen in the first split point of the trees about 92% of the time, the percentage of H atoms, the total path count, and the frequency of pairs of electronegative atoms such as O, Br, Cl, I, and F at certain topological distances. Many model-prioritized Dragon7 features for cruzain related to the frequency of pairs of electronegative atoms such as O, F, Cl, N, P, and S. Some top Dragon7 features which overlapped between the two models related to the frequency or existence of pairs of carbon atoms at certain topological distances. Based on visualizations of per-class distributions for these features, putative aggregators for only the β-lactamase set tend to have larger H%, while aggregators for both target sets tend to have significantly larger average molecular weight and total path counts. Model-prioritized ISIDA features for both targets were enriched in fragments containing many sp^3-hybridized carbon atoms. However, the only fragment which significantly distinguished the two classes was "C-H" for the β-lactamase model specifically, which tended to be significantly larger for the putative aggregator compounds. As each of these prioritized features broadly relate to weight, chemical complexity, aliphaticity, and polarity, we used simple RDKit features to determine the generalizability of these trends. Indeed, based on visualizations of class-separating distributions for these RDKit features, the putative aggregator class in both target datasets had a significantly higher frequency of heteroatoms and rings, although

only the β-lactamase dataset had an increase in the percentage of sp^3-hybridized carbons in the aggregator class. A series of other Dragon7 features significantly distinguished the two classes for both target datasets, such as HyWi_H2, AVS_Dt, ATS7p, and Log_{10}(Wap), although these features are more difficult to interpret and generalize. As the Aggregator Advisor[25] uses logP as a metric to flag aggregation, we evaluated the importance of the MLOGP Dragon7 feature and found that it was never selected at the first branch point in our models and was seldom chosen at the second or third points. Additionally, we demonstrate that MLOG does not significantly separate the putative aggregator and non-aggregator classes in either target dataset. Although there is a slightly larger trend for aggregators to have larger MLOG based on median and mean, there were considerable populations of putative aggregators which have MLOGP below 3.0 and non-aggregators with MLOG above 3.0. Although we posit that aggregators tend to be larger in molecular weight and more complex, and specifically for the β-lactamase set, also seem to have more aliphatic groups, we also discovered that AMW, FractionCSP3, and SlogP alone are insufficient to approximate the predictive accuracy attained by our consensus models on the β-lactamase curated set. For example, after optimizing threshold values to maximize classification accuracy on the entire unbalanced β-lactamase set, we found that SlopP at a threshold of 3.175 and FractionCSP3 at a threshold of 0.3175 alone obtain a balanced accuracy of about 0.60, while AMW at a threshold of 385 alone achieves a balanced accuracy greater than 0.69, approximately the same as when using all three classifiers together. In a fair comparison between our consensus models and RF models created from the AMW, SlogP, FractionCSP3, and C-H features on the balanced β-lactamase dataset we found that our models have more than a 10% improvement in 5FCV balanced accuracy and external non-aggregator set SP over the refined feature model, although this comes at the cost of applying an AD. Overall, the simplest small molecule attribute

which most accurately predicts aggregation is molecular weight. This makes sense, as solubility tends to decrease with increasing molecular size.

We encourage the application of our models by the scientific community. In the simplest form, our curated datasets for both the β-lactamase and cruzain targets provided in **Supplemental Material 2** could be used to identify compounds which have displayed detergent-sensitive activity. Screening libraries through both of our β-lactamase and cruzain models can help inform whether compounds should be orthogonally screened as aggregators or triaged based on specific HTS assay conditions. We recognize that our models may not be extensible to the prediction of aggregators in extremely variable assay conditions. We hope that these models increase public confidence and usage in computational approaches which predict specific mechanisms of assay interference and emphasize correct controls for particularly likely interference compounds. While these tools are economically helpful for analyzing massive libraries, we do not in any way advocate their usage as sole substitutes for well-designed experimental counter-screens for aggregation. We anticipate that the application of our predictors will inform medicinal chemists and chemical biologists of misleading aggregators and reduce wasted resources in pursuit of deceiving, non-tractable chemical leads.

5. Dissemination

These models have been encapsulated within a KNIME prediction workflow for VS and are accessible for download in **Supplemental Materials 3-7**. To use the workflow, import the Supplemental_Material_3.knwf to your KNIME working environment, input your Dragon7 and ISIDA features sets, point the "List Files" nodes to the directories containing the appropriate downloaded and extracted Dragon7 and ISIDA models for β-lactamase and cruzain, and execute the "Predictions" metanode. It has been assumed that the user has curated their screening data

per best practices, calculated the Dragon7 and ISIDA feature sets as previously described, set the headers of the unique compound identifier column to “Compound_name”, and maintained the feature headers names as given exactly by the software packages. Separate predictions for the β-lactamase and cruzain targets will be saved to a specified location in a csv file following prediction of the query dataset. These models can be employed by users to predict putative aggregators and non-aggregators in chemical libraries of their interest.

APPENDIX

Supplemental Material 1 - Real-time video demonstration of various examples of the Aggregator Advisor either successfully detected or failing to detect compounds within its own Rogues' Gallery

Supplemental Material 2 - Curated SDF balanced and unbalanced training sets and balancing-withheld non-aggregator and detergent-resistant test sets for the β-lactamase and cruzain targets

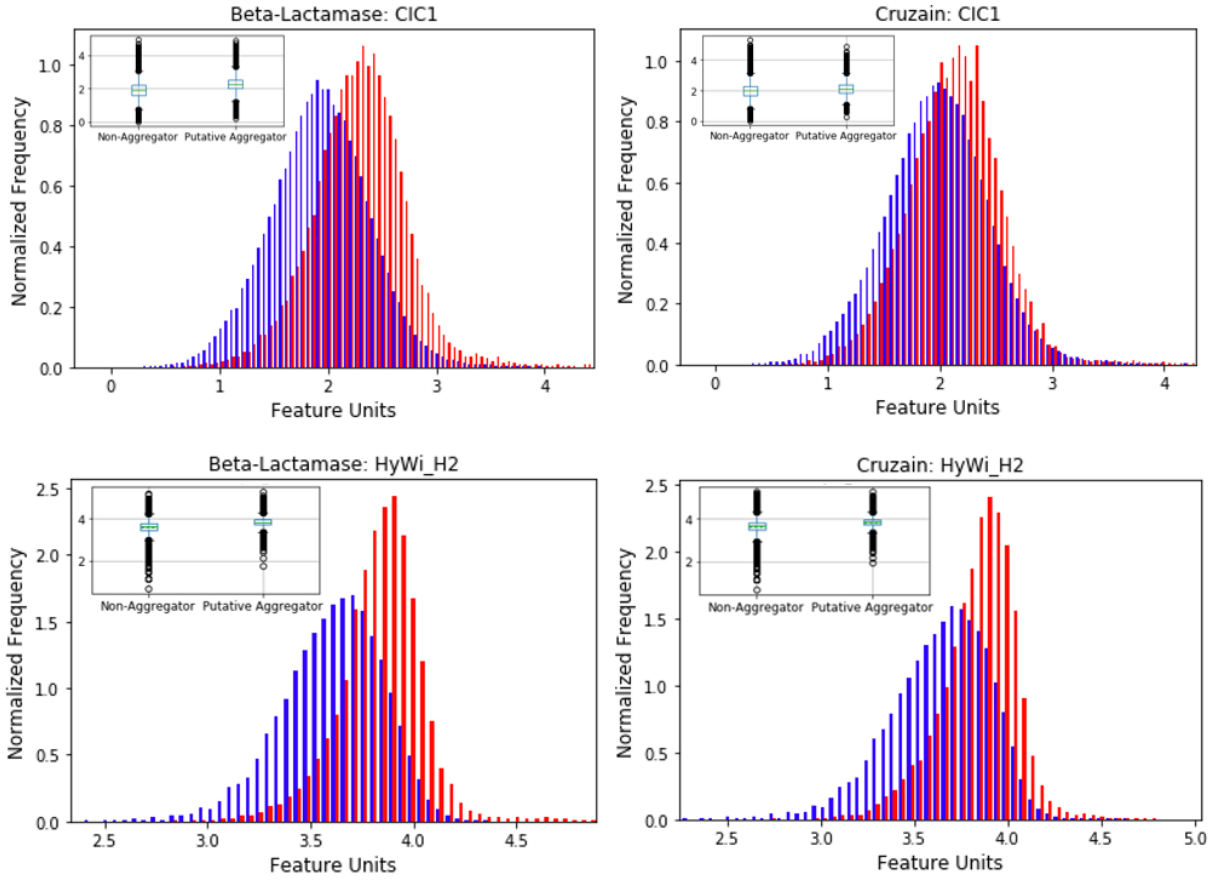
Supplemental Material 3 - KNIME prediction workflow

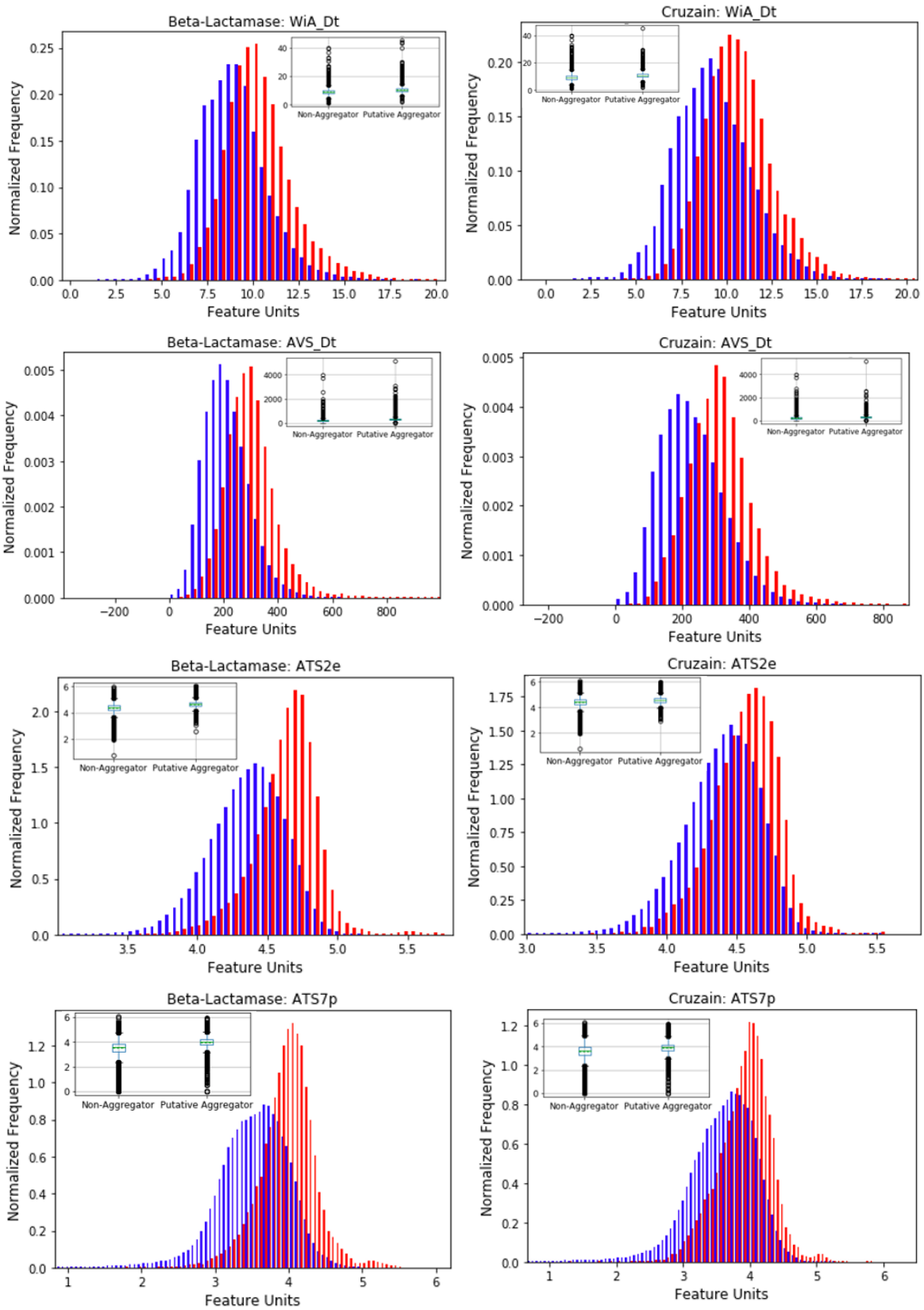
Supplemental Material 4 - β-lactamase Dragon7 models

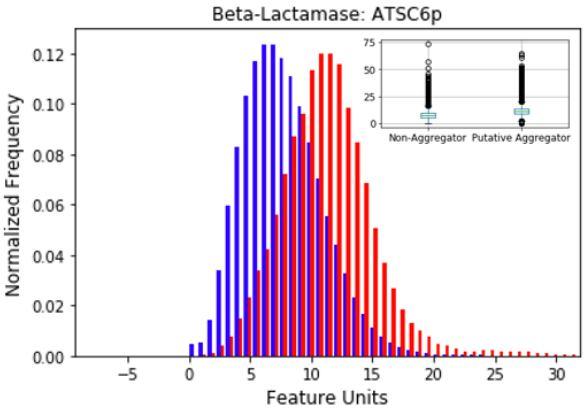
Supplemental Material 5 - β-lactamase ISIDA models and feature headers

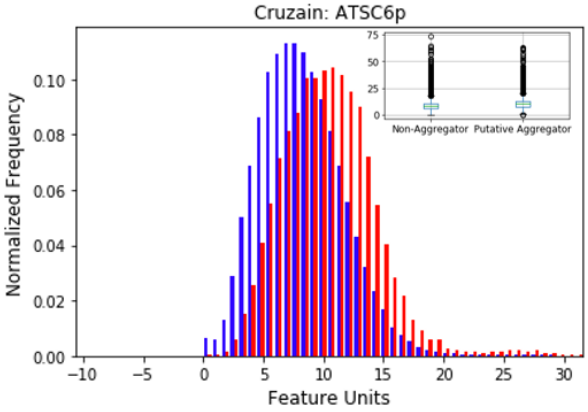
Supplemental Material 6 - Cruzain Dragon7 models

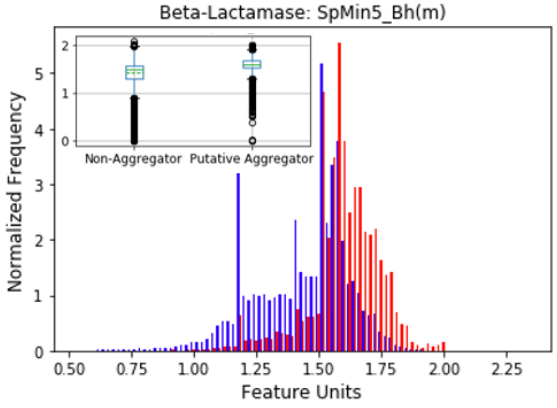
Supplemental Material 7 - Cruzain ISIDA models and feature headers

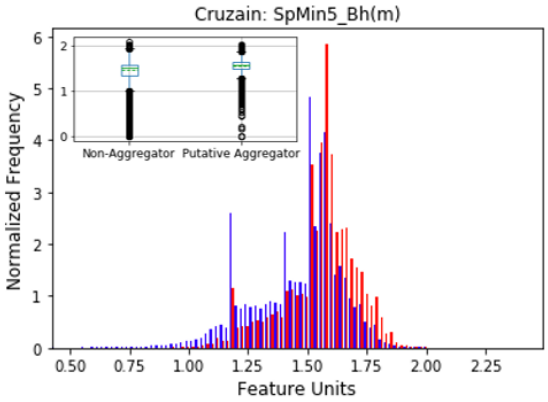


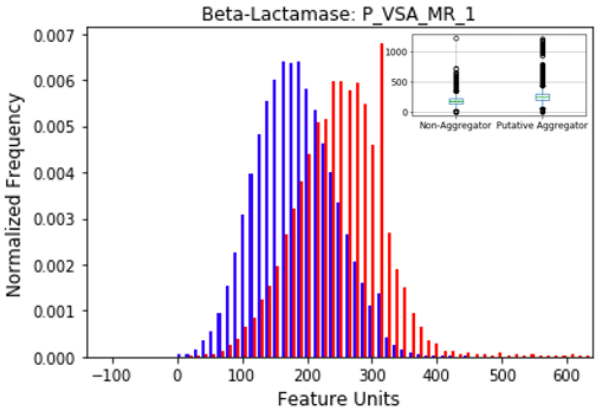

Beta-Lactamase: WiA_Dt
Cruzain: WiA_Dt
Beta-Lactamase: AVS_Dt
Cruzain: AVS_Dt
Beta-Lactamase: ATS2e
Cruzain: ATS2e
Beta-Lactamase: ATS7p
Cruzain: ATS7p
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator

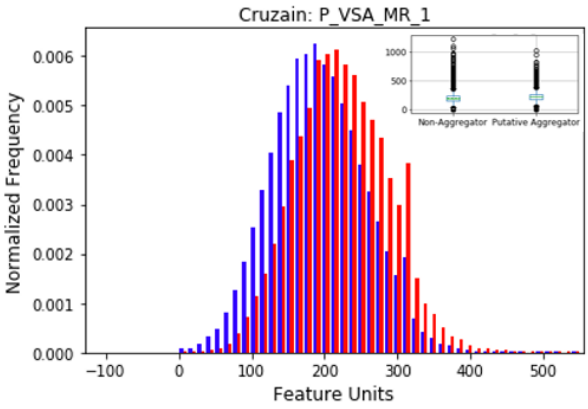

Beta-Lactamase: ATSC6p
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator

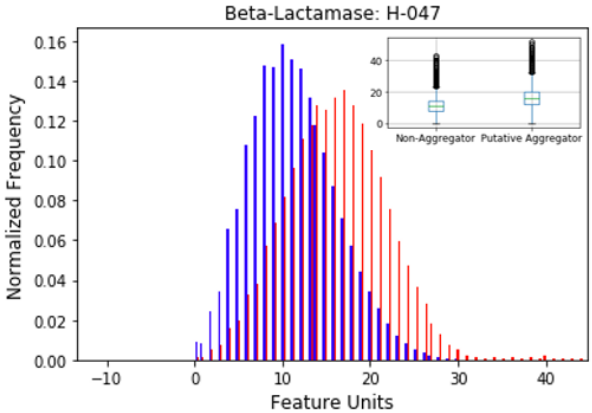

Cruzain: ATSC6p
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator

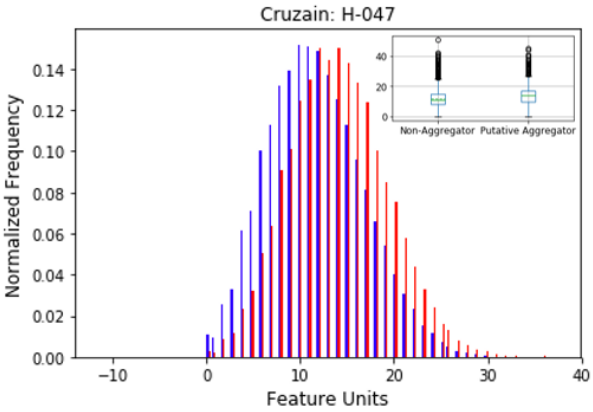

Beta-Lactamase: SpMin5_Bh(m)
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator

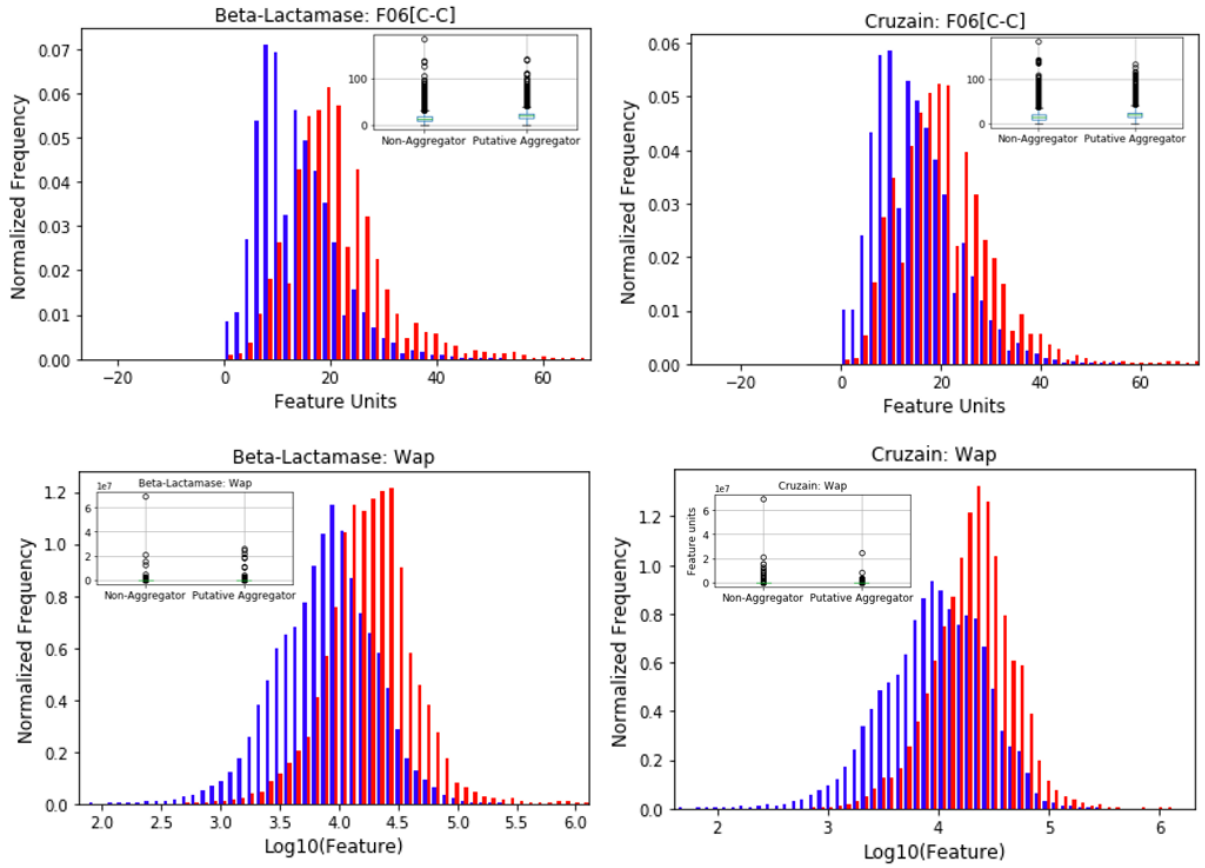

Cruzain: SpMin5_Bh(m)
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator


Beta-Lactamase: P_VSA_MR_1
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator


Cruzain: P_VSA_MR_1
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator


Beta-Lactamase: H-047
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator


Cruzain: H-047
Normalized Frequency
Feature Units
Non-Aggregator
Putative Aggregator



Supplemental Material 8 - Per-class distributions for additional select Dragon7 model-prioritized features

Histograms and inset box plots illustrating distributions for the non-aggregator (blue) and putative aggregator (red) classes on Dragon7 features selected based on their strong class-separating distributions from the entire unbalanced, curated β-lactamase (left column) and cruzain (right column) datasets. Histograms are normalized to possess a total area of 1. In most cases 100 bins were used. Note that the histogram for the last feature, Wap, was scaled to $\log_{10}$.

REFERENCES

(1) Martis, E. High-Throughput Screening: The Hits and Leads of Drug Discovery- An Overview | Elvis Martis - Academia.Edu. *J. Appl. Pharm. Sci.* **2011**, *01* (01), 02-10.

(2) Roy, A. Early Probe and Drug Discovery in Academia: A Minireview. *High-Throughput* **2018**, *7* (1), 4.

(3) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* **2007**, *50* (10), 2385–2390.

(4) Thorne, N.; Auld, D. S.; Inglese, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr Opin Chem Biol* **2010**, *14* (3), 315–324.

(5) Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *8* (3), 387–390.

(6) Dahlin, J. L.; Baell, J.; Walters, M. A. Assay Interference by Chemical Reactivity. *Assay Guid. Man.* **2004**, 1–43.

(7) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* **2015**, *58* (5), 2091–2113.

(8) Erlanson, D. A. Learning from PAINful Lessons. *J. Med. Chem.* **2015**, *58* (5), 2088–2090.

(9) Baell, J. B. Screening-Based Translation of Public Research Encounters Painful Problems. *ACS Med. Chem. Lett.* **2015**, *6* (3), 229–234.

(10) Duan, D.; Torosyan, H.; Elnatan, D.; McLaughlin, C. K.; Logie, J.; Shoichet, M. S.; Agard, D. A.; Shoichet, B. K. Internal Structure and Preferential Protein Binding of Colloidal Aggregates. *ACS Chem. Biol.* **2017**, *12* (1), 282–290.

(11) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53* (1), 37–51.

(12) Auld, D. S.; Inglese, J.; Dahlin, J. L. Assay Interference by Aggregation. *Assay Guid. Man.* **2017**, 1–33.

(13) Coan, K. E. D.; Shoichet, B. K. Stoichiometry and Physical Chemistry of Promiscuous Aggregate-Based Inhibitors. *J. Am. Chem. Soc.* **2008**, *130* (29), 9606–9612.

(14) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A Specific Mechanism of

Nonspecific Inhibition. *J. Med. Chem.* **2003**, *46* (20), 4265–4272.

(15) McLaughlin, C. K.; Duan, D.; Ganesh, A. N.; Torosyan, H.; Shoichet, B. K.; Shoichet, M. S. Stable Colloidal Drug Aggregates Catch and Release Active Enzymes. *ACS Chem. Biol.* **2016**, *11* (4), 992–1000.

(16) Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K.; Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding. **2009**, *52* (March), 2067–2075.

(17) Sassano, M. F.; Doak, A. K.; Roth, B. L.; Shoichet, B. K. Colloidal Aggregation Causes Inhibition of G Protein-Coupled Receptors. *J. Med. Chem.* **2013**, *56* (6), 2406–2414.

(18) 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.* **2002**, *45* (8), 1712–1722.

(19) 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.* **2008**, *51* (3), 574–580.

(20) Ryan, A. J.; Gray, N. M.; Lowe, P. N.; Chung, C. Effect of Detergent on "Promiscuous" Inhibitors. *J. Med. Chem.* **2003**, *46*, 3448–3451.

(21) Babaoglu, K.; Simconov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; et al. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against β-Lactamase. *J. Med. Chem.* **2008**, *51* (8), 2502–2511.

(22) Mott, B. T.; Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Ang, K. K. H.; Leister, W.; Shen, M.; Silveira, J. T.; Doyle, P. S.; Arkin, M. R.; et al. Identification and Optimization of Inhibitors of Trypanosomal Cysteine Proteases: Cruzain, Rhodesain, and TbCatB. *J. Med. Chem.* **2010**, *53* (1), 52–60.

(23) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-Throughput Assays for Promiscuous Inhibitors. *Nat. Chem. Biol.* **2005**, *1* (3), 146–148.

(24) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46* (21), 4477–4486.

(25) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58* (17), 7076–7087.

(26) Rao, H.; Li, Z.; Li, X.; Ma, X.; Ung, C.; Li, H.; Liu, X.; Chen, Y. Identification of Small Molecule Aggregators from Large Compound Libraries by Support Vector Machines. *J.*

Comput. Chem. **2010**, *31*, 752–763.

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

(28) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20* (3–4), 241–266.

(29) Stork, C.; Wagner, J.; Friedrich, N. O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13* (6), 564–571.

(30) Stork, C.; Wagner, J.; Friedrich, N. O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter 2.0: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13* (6), 564–571.

(31) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.

(32) Tanimoto, T. *IBM Internal Report. Armonk: IBM Corp*; 1957.

(33) Shoichet, B. K. Interpreting Steep Dose-Response Curves in Early Inhibitor Discovery. *J. Med. Chem.* **2006**, *49* (25), 7274–7277.

(34) Southall, N. T.; Jadhav, A.; Huang, R.; Nguyen, T.; Wang, Y. *Handbook of Drug Screening*, Second.; Seethala, R., Zhang, L., Eds.; informa healthcare: New York, New York, USA, 2009.

(35) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11*, 535.

(36) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(37) Fourches, D.; Muratov, E. N.; Tropsha, A. Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, No. 56, 1243–1252.

(38) Kode. DRAGON7.0 https://chm.kode-solutions.net/products_dragon.php (accessed October 9, 2016).

(39) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided. Drug Des.* **2008**, *4* (3), 191–198.

(40) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J.*

Chem. Inf. Model. **2014**, *54* (1), 1–4.

(41) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(42) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22* (1), 69–77.

(43) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Random Forests. In *An Introduction to Statistical Learning*; Springer-Verlag New York: New York, New York, USA, 2013; pp 320–321.

(44) Breiman, L. E. O. Random Forest. **2001**, 5–32.

(45) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (1), 18–28.

(46) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932–2942.

(47) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53* (13), 4891–4905.

(48) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54* (3), 705–712.