Detection of frequent hitters and a study of their frequency across various screening technologies

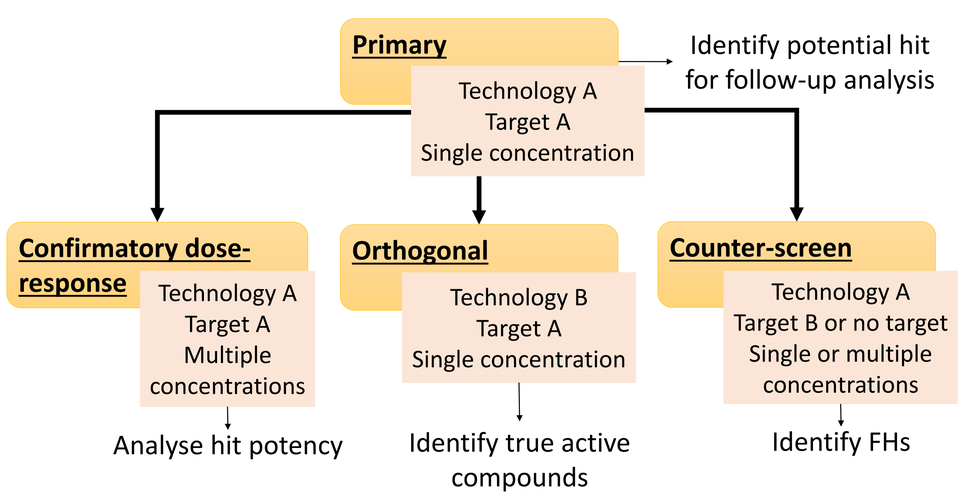
Laurianne David[†‡], J.W.M. Nissink[ξ], J. Walsh[ϫ], I. Feierberg[ϟ], P. Várkonyi[†], J. Bajorath[‡] , H. Chen[†], O. Engkvist[†]

[†] Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden, [ξ] Chemistry, Oncology, IMED Biotech Unit, AstraZeneca, Cambridge, CB4 OWG, UK, [ϫ] Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, UK, [ϟ] Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Boston, US, [‡] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

- A known **issue in high throughput screening** (HTS) is the presence of **frequent hitters**[1] (FHs), because they impede research and waste time and resources, by being investigated in follow-up studies at the expense of other experiments. FHs include problematic (e.g., reactive, impure or aggregating) compounds, promiscuous compounds that exhibit the desired mode of action on the target but are non-selective across targets, and compounds that are false actives across many assays, due to interference with the assay technology. Such interference may arise through e.g., inhibition of a coupled enzyme in the assay or fluorescence quenching properties.
- Herein we present a comparison of two predictive models designed to **identify frequent hitters**. These models are applied to various HTS technologies (FRET, ADP-Glo™ [2]…) data. Their advantages, disadvantages and predictive powers are discussed. The first model[3] relies on historical HTS activity data, the second on structural properties.

Data Collection

- Screening data related to the ADP-Glo™ and FRET technologies are assembled from the AstraZeneca screen collection.
- In primary assays, compounds' activity flags (active and inactive) were analyzed to assign the most accurate flag to compounds. Indeed, compounds can occasionally have two different activity flags in the same assay. This may be due to the use of different assay plates, a solubility error or the test being done at two concentrations. In order to overcome this discrepancy, a number of rules were applied (the molecular weight and the concentration of the compounds were considered as well as the most recurrent activity flag found in primary and confirmatory assay). For example, if a high molecular weight compound (> 350 g/mol) is tested at 10 and 100 µM, the activity flag at 10 µM is considered to be the most accurate to overcome the possibility of the flag being biased by an aggregation.
- For each technology, the newest and most relevant assay cascade containing a counter-screen assay was identified. Compounds were predicted as being FHs or not based on data from the curated primary assays and their corresponding orthogonal assays. Predictions were then validated using the data available in the counter-screen assays (i.e. a compound active in a counter-screen assay is a FH).



Prediction of anomalous behavior using machine-learning and historical data

Random Forest Classifier (RF)

	Compound IDs	Activity in HTS assay	Activity in orthogonal assay	Probable behaviour	ECFP4
Training set	XX0	A	A	Non-FH (i.e. true active)	0001001
	XX1	A	I	FH (i.e. false active)	1110110
Compound to predict	XX2	?	?	?	0010110
	XX3	?	?	?	0101011

Compound	Predicted behaviour
XX2	FH
XX3	Non-FH

Model parameters
- Number of trees in the forest = 150
- Class weight = balanced
- Maximum number of features considered by each tree = sqrt(number of features)
- Minimum number of samples to be a leaf node = 5

Binomial Survivor Function (BSF)

Compound IDs	Technology	# Time tested (N)	# Time active (a)	Hit Rate over all assays (h)	Target class	# Time tested (N)	# Time active (a)	Hit Rate over all assays (h)
XX0 **	**ADP-Glo™**	12	12	0.01	**Kinase**	20	19	0.01
XX1 **		3	2			18	2	
XX2		14	1			0	0	

*** Compound tested on a unique target class*

$$pBSF = -\log \sum_{a=A}^{N} \binom{N}{a} h^a (1-h)^{N-a}$$

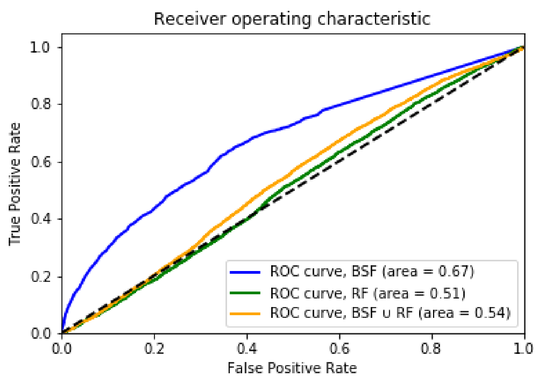
Is FH when pBSF > 2

Compound IDs	pBSF (technology)	pBSF (target class)	Is FH (technology)
XX0 **	24	36.7	Unsure
XX1 **	3.5	1.8	Yes
XX2	0.88	0	No

Anomalous behavior in assay technology

➢ FRET technology

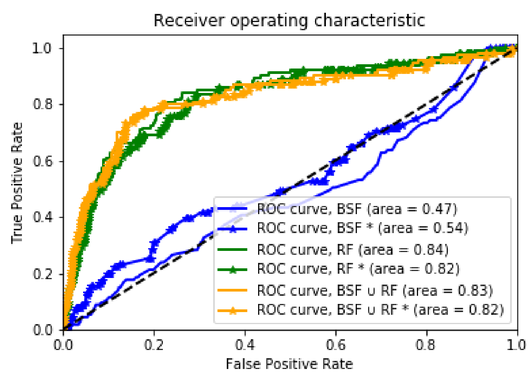
The RF model performs slightly better at predicting FHs than the BSF, however the later is more efficient at correctly predicting non-FH compounds. Indeed, BSF and RF correctly identified, respectively, 1,332 and 2,060 out of 2,977 FHs as well as 10,844 and 4,601 out of 13,720 non-FH compounds.



	BSF	RF	BSF ∪ RF
Recall	0.45	0.69	0.83
Precision	0.32	0.18	0.20
Specificity	0.79	0.34	0.27

➢ ADP-Glo™ technology

Most of the ADP-Glo™ assays were tested on a kinase. A BSF calculation based on these assays could classify kinase promiscuous compounds as ADP-Glo™ FHs. Therefore, BSF was calculated on all the ADP-Glo™ assays, as well as on all the other available kinase assays. BSF* and RF* correctly identified, respectively, 59 and 78 out of 113 FHs as well as 9,445 and 11,446 out of 13,943 non-FH compounds.



	BSF	BSF*	RF	RF*	BSF ∪ RF	BSF ∪ RF*
Recall	0.67	0.52	0.84	0.69	0.97	0.78
Precision	0.01	0.01	0.02	0.03	0.01	0.01
Specificity	0.26	0.68	0.73	0.82	0.09	0.55

** Kinase promiscuous compounds are classified as non-FHs. The models show then a decrease in recall but a significant increase in specificity.*

Discussion

- The **BSF model** relies on **historical activity data** to predict compounds behaviours and thus requires the **compounds to have been tested at least once**. The impact of target classes should always be considered to filter out promiscuous compounds from predicted frequent hitters.
- The **RF model** relies on **structural information** and, contrary to the BSF model, can be applied to **compounds that were never tested before**.
- Both predictive models show a low precision for the ADP-Glo™ technology. With the RF model, this can be improved by better data extraction to build a stronger training set. With the BSF, this is expected to improve with widespread adoption of the technology and new data becoming available for integration.
- BSF can be applied when the identification of as many true active compounds as possible is as important as the identification of FHs. RF can be of use when the aim of a project requires the detection of as many FHs as possible (i.e. when the number of compounds to test is limited and when it is necessary to be sure to have only true hits, e.g. for financial reasons).
- **Other assay cascades** will be investigated to derive a general analysis over each technology.
- For future work, emphasis will be placed on having **better data extraction**, and on the usage of new **structural fingerprints** to build machine-learning models (e.g.§, physico-chemical fingerprints or scaffold analysis).

References

1. O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, EM. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, J. Med. Chem., **2002**, 45 (1), 137-142.
2. H. Zegzouti, M. Zdanovskaia, K. Hsiao, SA. Goueli, Assay Drug Dev. Technol., **2009**, 7 (6), 560-72.
3. J.W.M. Nissink, S. Blackburn, Future Med. Chem., **2014**, 6 (10), 1113-26.

This work was supported financially by the **European Union's Horizon 2020 research and innovation program** under the Marie Skłodowska-Curie grant agreement No. 676434, "Big Data in Chemistry" (**"BIGCHEM",** http://bigchem.eu).