

Jeremy J. Yang^{1,2}, Oleg Ursu^{1,2}, Christopher A. Lipinski³, Larry A. Sklar², Tudor I. Oprea^{1,2} and Cristian G. Bologa^{1,2}

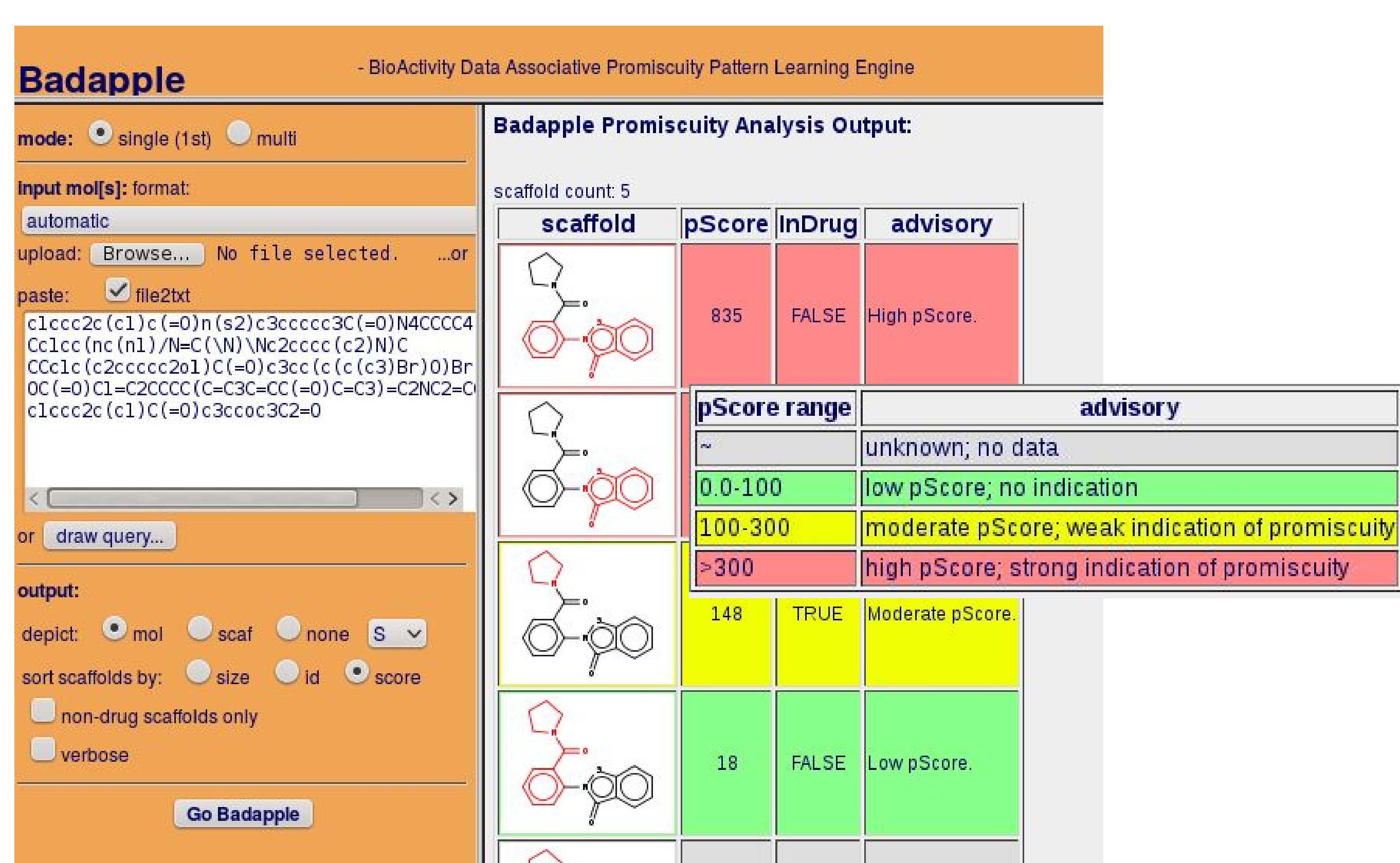
¹Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA

²Center for Molecular Discovery, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA.

³10 Connshire Drive, Waterford, CT 06385-4122, USA

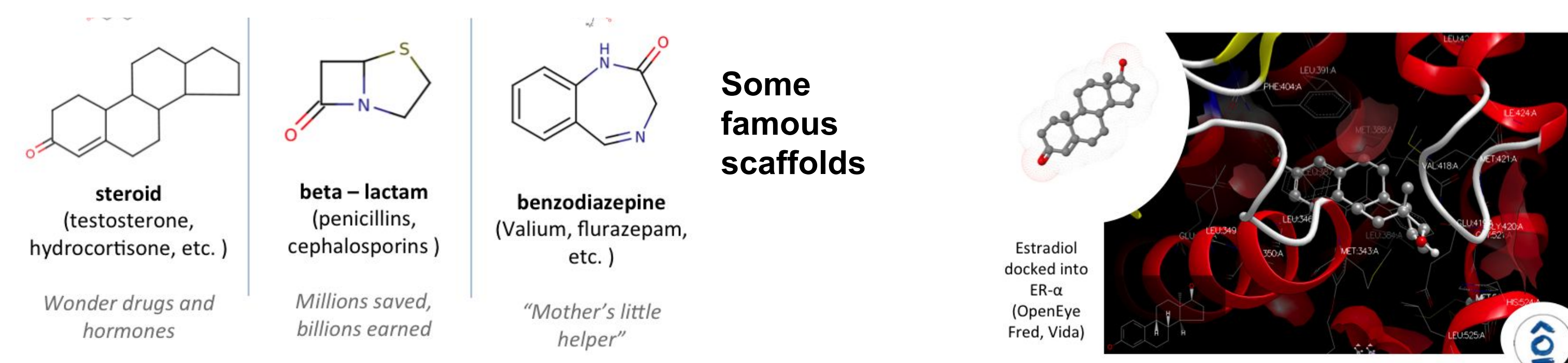
Bioassay hit selection; stacking the odds in early stage drug discovery

High throughput screening (HTS) data analysis continues to be an essential, routine, yet challenging task in drug discovery: to infer reliable knowledge from big and noisy data. Bioassays require complex methodology, and results vary widely in accuracy, precision, and content. Hit selection criteria should optimize the overall probability of success in a project, and avoid expensive “false trails” such as promiscuous compounds. At UNMCMD, our experience in the NIH Molecular Libraries Project (MLP) motivated and informed this research.



Badapple statistical learning

Badapple (bioassay data associative promiscuity prediction learning engine) is an algorithm, software system, and online service for identifying likely promiscuous compounds via associated scaffolds to assist and accelerate drug discovery informatics. Predictions are based on analysis of empirical data from NIH MLP assays. Badapple has been released via: (1) BARD REST API and web client, and (2) public web app.



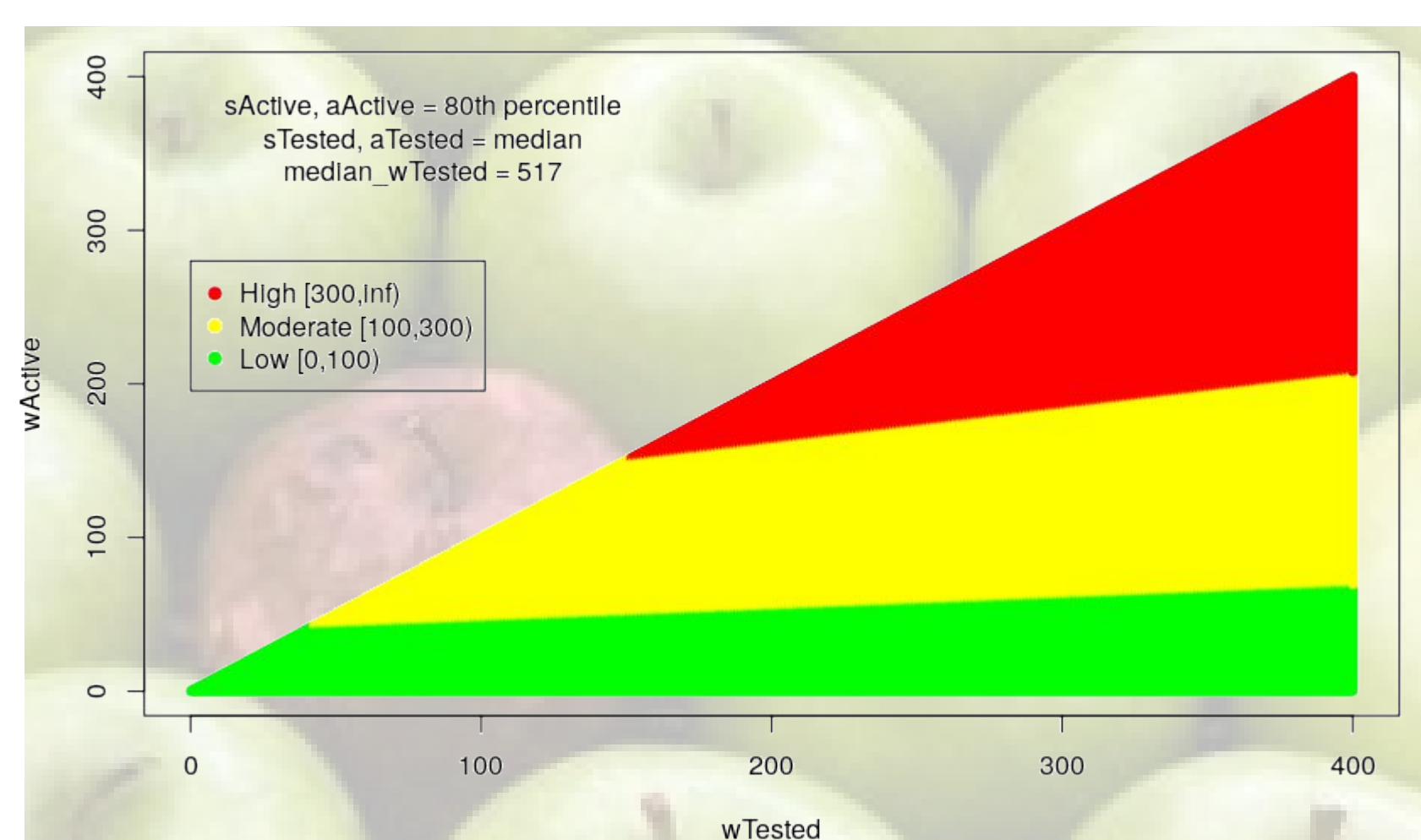
Why scaffolds?

Scaffolds are useful for several reasons: (1) Scaffolds relate analog chemical series, relevant to medicinal chemistry and lead optimization. (2) Data may not exist about a specific compound, but may exist about a closely related compound with common scaffold. (3) “Privileged structures” theory suggests scaffolds often confer bioactivity, via 3D shape or binding interactions. Badapple employs the HierS algorithm, and complements other methods, such as “Pan-Assay INterference CompoundS” (PAINS) which features expert curation of substructure patterns. In contrast, Badapple is fully automatic, and fully empirical.

Promiscuity defined pragmatically

For simplicity, comprehensibility, and practical utility, “promiscuity” is defined as multiplicity of positive non-duplicate bioassay results -- i.e. target multiplicity. It is well understood that positives (hits) may be false, due to experimental artifact (e.g. aggregation, reactivity, fluorescence). Yet, such a compound will generally be undesirable regardless.

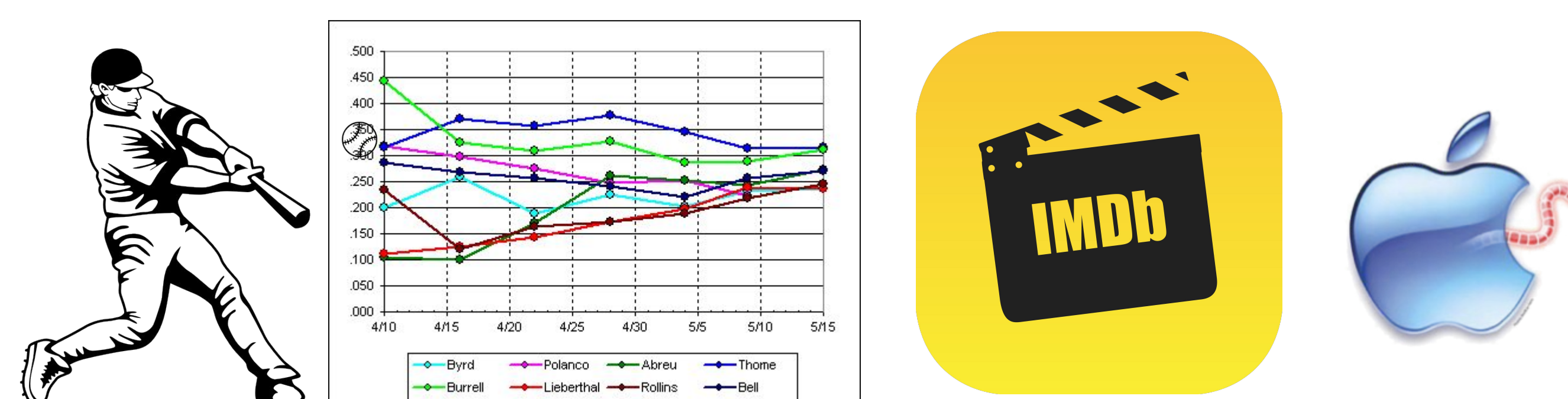
$$score = \frac{s_A}{s_T + med(s_T)} * \frac{a_A}{a_T + med(a_T)} * \frac{w_A}{w_T + med(w_T)} * 1e5$$



sT = tested substances
sA = active substances
aT = assays with tested compounds
aA = assays with active compounds
wT = tested samples
wA = active samples
all counts: for given scaffold

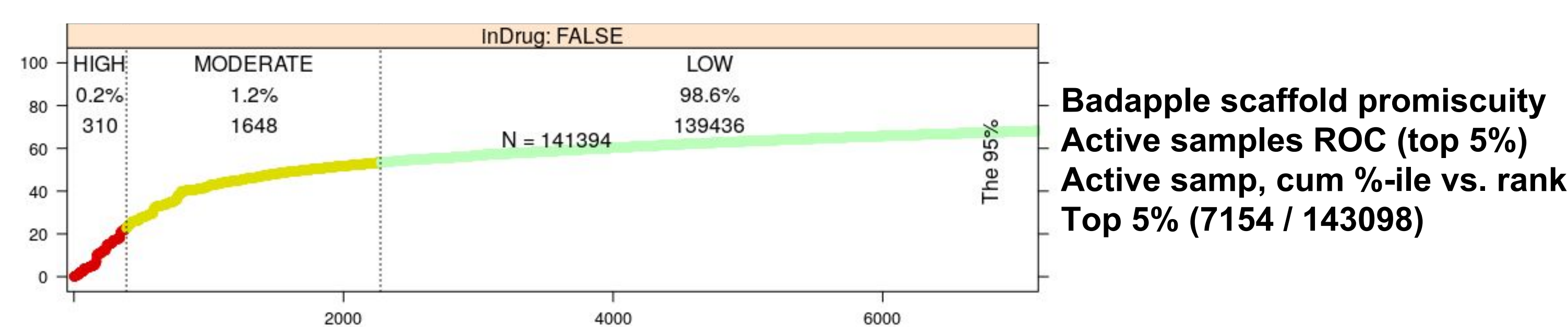
The Badapple formula

The Badapple promiscuity score is a product of three terms, related to substances, assays and samples, each important to produce a high score. Global medians normalize scores to reflect weight of evidence and statistical learning.



Baseball, movies, and strong evidence

As with baseball hits, there is randomness and noise in bioassay data, and sufficient sampling is essential. Early season batting averages (BA) are not predictive, but converge over time. The Badapple formula also shares some properties with the IMDb score used to rank movies, which considers vote count for weight of evidence.



Results: the privileged and notorious few

Although there are relatively few high scoring scaffolds, those “privileged” few account for a disproportionate share of the bioactivity. Overall, 50% of all bioactivity is associated with 1.4% of the scaffolds. The medicinal chemist among us (CAL) has identified mechanisms of promiscuity validating several top scoring scaffolds.

Conclusion

Badapple can identify “false trails” and streamline bioassay workflows, improving the odds in early stage drug discovery.

References:

1. NIH Molecular Libraries Program (2004-2013). <https://commonfund.nih.gov/molecular-libraries>.
2. Wilkens SJ, Janes J, Su AI (2005) HierS: hierarchical scaffold clustering using topological chemical graphs. J Med Chem 48:3182–3193
3. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740.
4. BioAssay Research Database (BARD) (2012). <http://bard.nih.gov>.
5. IMDb Votes/Ratings Top Frequently Asked Questions. http://www.imdb.com/help/show_leaf?votestopfaq.
6. JChem 5.8.3, ChemAxon (2012). <http://www.chemaxon.com>.
7. Badapple public web app (2013). <http://pasilla.health.unm.edu/tomcat/badapple>.
8. Badapple: promiscuity patterns from noisy evidence, Yang JJ, Ursu O, Lipinski CA, Sklar LA, Oprea TI Bologa CG, J. Cheminfo. 8:29 (2016), DOI: 10.1186/s13321-016-0137-3.