

Managing missing measurements in small-molecule screens

Michael R. Browning · Bradley T. Calhoun ·
S. Joshua. Swamidass

Received: 19 January 2013 / Accepted: 29 March 2013 / Published online: 13 April 2013
© Springer Science+Business Media Dordrecht 2013

Abstract In a typical high-throughput screening (HTS) campaign, less than 1 % of the small-molecule library is characterized by confirmatory experiments. As much as 99 % of the library's molecules are set aside—and not included in downstream analysis—although some of these molecules would prove active were they sent for confirmatory testing. These missing experimental measurements prevent active molecules from being identified by screeners. In this study, we propose managing missing measurements using imputation—a powerful technique from the machine learning community—to fill in accurate guesses where measurements are missing. We then use these imputed measurements to construct an imputed visualization of HTS results, based on the scaffold tree visualization from the literature. This imputed visualization identifies almost all groups of active molecules from a HTS, even those that would otherwise be missed. We validate our methodology by simulating HTS experiments using the data from eight quantitative HTS campaigns, and the implications for drug discovery are discussed. In particular, this method can rapidly and economically identify novel active molecules, each of which could have novel function in either binding or selectivity in addition to representing new intellectual property.

Introduction

In recent years, small-molecule high-throughput screening (HTS) technology has dramatically impacted research in both the pharmaceutical industry and academics [1], and is an increasingly common method of identifying molecules with similar properties. High-throughput screening projects are typically structured in multiple stages. In the first stage, a primary screen is executed at a single dose. In the second stage, a small subset of hits from the first stage are verified by a more accurate “dose-response” experiment to determine potency (Fig. 1). Hits confirmed in dose-response are then sent for further computational and experimental analysis while, usually, the primary screen is ignored from here forward.

A fundamental issue in these campaigns is that only a small fraction of molecules are fully characterized by confirmatory and follow-up experiments. Therefore, rather than comprehensively surveying an entire molecule library, as much as 99 % of the library is incompletely characterized. The majority of these molecules are not active and not worth the effort of characterizing. However, some of them may prove active with further testing [2–6], meaning that the missing confirmatory experiment measurements for these molecules are important.

Three strategies have been used to manage the uncertainty introduced by this missing data. In the first strategy, screeners capture potency data for all molecules in the library, in a protocol called quantitative high-throughput screening (qHTS) [7]. Clearly, qHTS generates higher quality data and identifies virtually all of the active molecules in the library. However, most HTS campaigns do not use qHTS methodology. Moreover, qHTS does not entirely remove the problem of missing measurements; even if the primary screen is a qHTS, subsequent confirmatory and

M. R. Browning · B. T. Calhoun · S. Joshua. Swamidass (✉)
Division of Laboratory and Genomic Medicine, Department of
Pathology and Immunology, Washington University School of
Medicine, St. Louis, MO, USA
e-mail: swamidass@wustl.edu

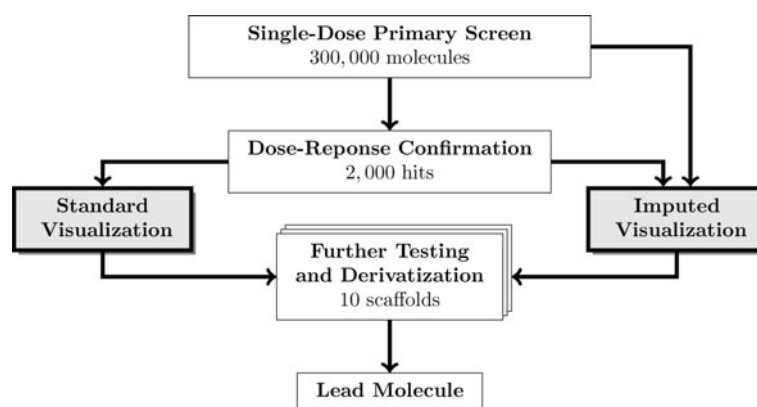


Fig. 1 Design of a typical HTS project. In this hypothetical example with typical numbers, a primary screen at a single dose is followed by a dose-response confirmatory test. Commonly, the results of the confirmatory experiments alone are collated to construct a visualization that chemists then use to select groups of molecules with

common scaffolds for further study. This study, in contrast, uses information from both the primary and confirmatory assays to construct an imputed visualization to more effectively choose scaffolds for further study

follow-up assays are executed on only a small fraction of molecules. In this case, the problems associated with missing measurements shift one step down the pipeline rather than being fully resolved.

In the second strategy, better hit-selection algorithms more carefully select molecules from the primary screen for follow-up, so as to ensure that the most interesting molecules are fully characterized [4–6, 8, 9]. These algorithms can work by leveraging molecular similarity [4], statistically modeling the results of confirmatory experiments [5], or modeling screener preferences [6]. These algorithms mitigate the impact of missing measurements by reducing the number of interesting but uncharacterized molecules, but they still leave as much as 99 % of the measurements missing.

In the final strategy, analytic methods *impute*—or predict using statistical models—the missing measurements. For several years, this strategy has been used extensively in the machine learning literature [10], and has recently been applied to sub-structure mining in HTS [11]. In HTS experiments, simple statistical models can label molecules not sent for confirmatory testing with the probability that they would be active if tested [5]. Effectively using these imputed measurements can dramatically improve the performance of sub-structure mining tools. This work parallels similar strategies recently proposed for large scale QSAR modeling [12].

In this study, we aim to further develop this final strategy by using imputation to improve the visualization of HTS results. We use statistical models to impute missing experimental measurements and map these measurements on to the scaffold tree visualization introduced by Schuffenhauer et al. [13]. Validating our method using several qHTS studies, we demonstrate that this visualization can reliably highlight active scaffolds missed by standard visualizations.

Data

We use the data from eight qHTS projects downloaded from PubChem (Table 1). Unlike typical HTS experiments, qHTS screens test all molecules at several doses and, therefore, assign an EC_{50} to every molecule in the project [7]. For our purposes, we consider molecules “active” if their inhibition potency is better than 5 μ M. Other thresholds yield similar results. As will be described, these qHTS screens are used to simulate the more commonly used HTS design: a single-dose primary screen followed by a dose-response confirmatory experiment.

Methods

Simulating multistage HTS

For each of our qHTS datasets, we simulate a single-dose primary screen by treating the measured activity at the dose closest to 10 μ M as the HTS score. We then simulate a dose-response confirmatory test on that data by extracting the potencies of the top 2,000 molecules with the best performance in our simulated primary screen, as well as the potencies of 1,000 molecules randomly selected from the screen. The EC_{50} potencies of those molecules not chosen are blinded and used only to validate our methods as a gold standard. Simulating a two-stage HTS workflow allows us to evaluate the efficacy of our methods by comparing predicted results on the blinded data against known potencies.

Imputing missing data

As will be seen, there is a clear relationship between a molecule’s measured activity in the primary screen and the

Table 1 qHTS datasets used in this study

	Size	Actives	PubChem ID	Inhibition target
ATA	330,119	4,174	504,466	ELG1 protein [16, 17]
BAZ	357,940	1,165	504,333	BAZ2B protein [18]
HLM	296,317	7,754	504,332	Histone lysine methyltransferase G9a [19, 20]
JMJ	368,479	2,996	504,339	JMJD2A-tudor domain [21]
MIR	333,521	3,282	2,289	miR-21 MiRNA [22, 23]
PME	388,720	4,729	588,591	POLH polymerase [24, 25]
TDP	290,745	707	485,290	Tyrosyl-DNA phosphodiesterase [26, 27]
TRX	38,6666	3,978	588,453	Thioredoxin reductase [28, 29]

Each data set is publicly available from the PubChem repository [14, 15] and the references in the table provide complete experimental details. The “Size” column is the total number of molecules tested, the “Actives” column is the total number of inhibitors identified with EC_{50} better than 5 μ M

probability it will be found to be an active in dose-response. The higher the activity in the primary screen, the more likely it is to be potent in the dose-response follow up. This relationship can be visualized by plotting the proportion of active molecules in subsequent batches. A key—and somewhat surprising—finding of our prior work is that data from the first few batches can be used to accurately guess the proportion of molecules active in subsequent batches [5, 11]. This enables us to label each untested molecule with the probability it would be found active if sent for confirmatory tests.

The basic idea is to fit a curve that maps the primary screening activity to the confirmation rate, and then use this curve to label the untested molecules. Many curve-fitting algorithms can be used find the relationship between activity and potency. We use a neural network with three hidden nodes (NN3) [30], using the screen activity as the single independent variable and the result from the dose-response experiment as the single dependent variable. We train the NN3 using gradient descent on the cross-entropy error, with a Gaussian prior on the weights [30]. By using a label of 1 for active and 0 for inactive (where active molecules have an EC_{50} of at least 5 μ M), the output of NN3 can be interpreted as the probability that a test molecule is active given its performance in the single-dose primary screen.

It is important to note that the data used to fit the curve is not a representative sample of the entire dataset; the curve is forced to extrapolate from the training set. The data from the first few batches is primarily made up of molecules with high activities in the screen. Therefore, sometimes imputed labels for those molecules with low activities are incorrect. We mitigate this source of error by using the curve to label only the most-active 18,000 untested molecules; we impute the remaining molecules as inactive.

Scaffold tree visualization

Scaffold trees organize the molecules in a screen into a tree, where each node is a scaffold encoded as a molecular framework. The framework assigned to each node in the tree fully contains the framework of its parent node. Each scaffold is associated with a set of molecules that are derivatives of its framework, and the activities of these molecules are used to color the scaffold’s node. We use the algorithm described by Schuffenhauer et al. [13] to generate each tree. The color of each node is computed from the number of active and inactive molecules associated with each node according to the following rules:

1. The hue of each node ranges from blue (inactive) to red (active) based on the fraction of associated molecules that are active.
2. The lightness of each node is scaled by the product of the number of active molecules with the fraction of active molecules, so that nodes with more active molecules are more vivid.
3. Nodes with no experimental data are colored white.

Colors usually range from light blue to dark red. Occasionally, a scaffold will be dark blue when a scaffold has large number of active molecules but also has a high number of inactive molecules. The computed coloring is dependent on the number of active and inactive molecules associated with each node, which is computed one of three ways.

Gold standard

Node colors are computed exactly using the qHTS data to produce a gold-standard tree against which to compare other methods.

Naive

In the naive strategy, only molecules sent for dose-response confirmation are considered. This simulates the typical HTS analysis, where all of the molecules below a certain activity threshold are removed from consideration.

Imputed

In the imputation strategy, all the molecules in the screen are considered. The missing measurements from the dose-response experiment are imputed. The number of active molecules is estimated by summing the imputed activities—or the experimental activity if known—of the molecules at each node.

Results

Accuracy of imputed labels

We use the NN3 to impute the activity of each molecule in the screen. We can predict the proportion of active molecules in an experiment, the confirmation rate, by averaging imputed activity across the set. We compare the predicted confirmation rate to the actual confirmation rate as determined by the qHTS study. Although there is some systematic error in the labels from the TDP study, the predicted confirmation rate and the actual confirmation rate are very similar across all datasets (Fig. 2).

Evaluation strategy

To evaluate our imputed visualization strategy, we compare the trees generated by the three different methods described above:

1. The *imputed* trees are computed by combining the unblinded potencies from the dose-response confirmatory test with the imputed potencies.
2. The *naive* trees are computed using only the molecules tested in the dose-response confirmatory test.
3. The *gold-standard* trees are computed using all the potencies from the qHTS screen.

In all three cases, molecules are considered active if their inhibitory potency is less than 5 μ M. To facilitate comparisons, throughout the paper, we display each node with its imputed color, but also include two switches to the right of each node with the naive (left) and gold-standard (right) color (Figs. 3, 5). We expect that the imputed trees will be closer than the naive trees to the gold standard. In actual use, the two switches would be removed (Fig. 4).

Accuracy of scaffold imputation

We test the accuracy of the imputation labels when grouped by scaffold by comparing the percentage of actives per scaffold as reported by the imputed and naive methods against the percentage of actives per scaffold as reported by the gold-standard method (Fig. 4). The average correlation between the imputed and gold-standard methods for all datasets is 0.77, with the highest correlation (from project TDP) being 0.87 and the lowest (from project

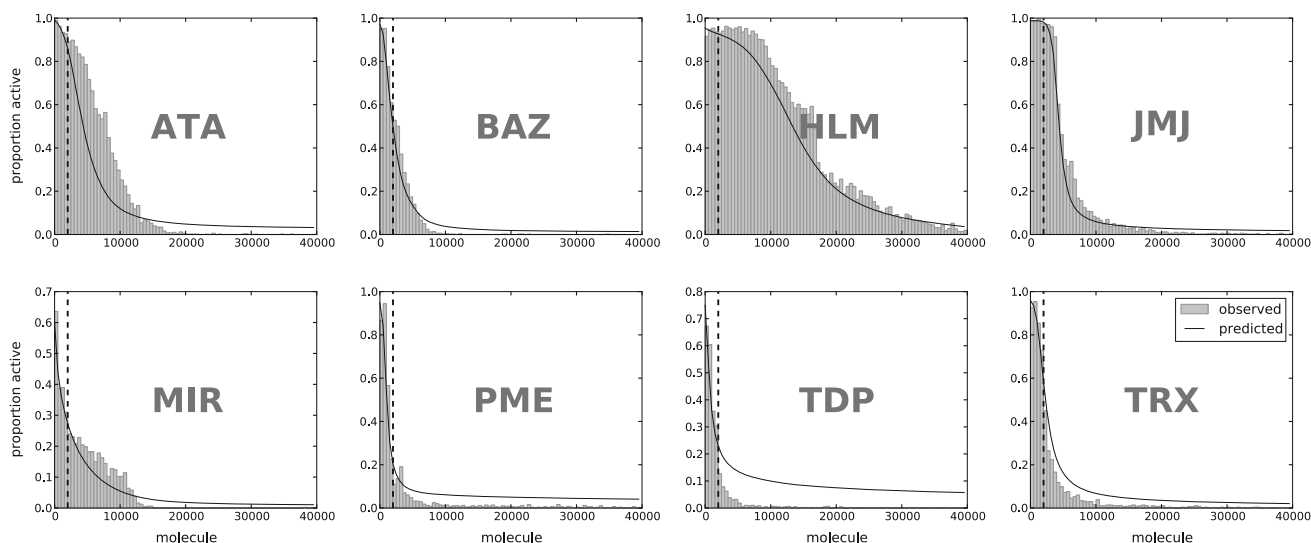


Fig. 2 Labeling accuracy. For each dataset, the best 40,000 molecules—ordered by their HTS activity—are grouped in batches of 500. The height of the bars corresponds to the observed confirmation rate

and the *solid line* corresponds to the average prediction of each batch. The bulk of the training molecules are to the *left* of the *dotted line* fixed at 2,000 molecules

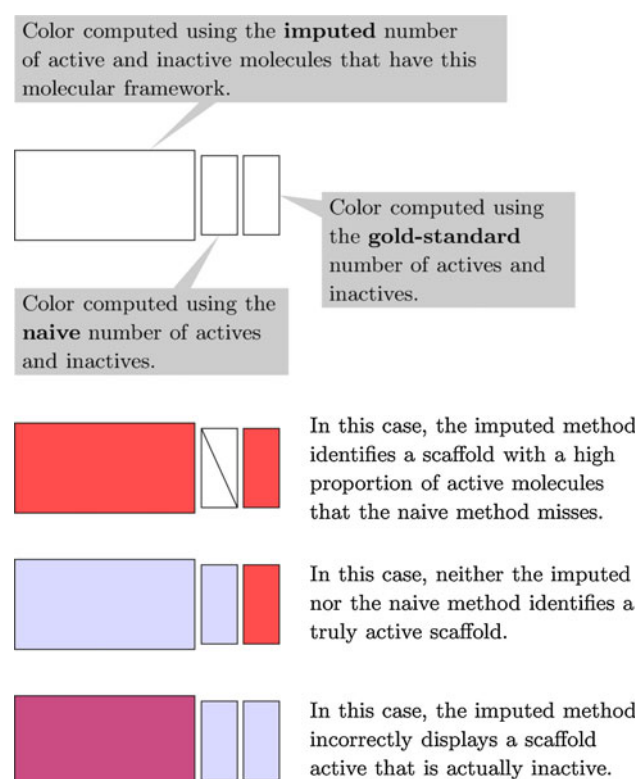


Fig. 3 Evaluation strategy. Throughout this paper, scaffolds are displayed using nodes like this figure. Each *panel* of the node displays a *different color* computed using either the imputed, naive, or gold-standard number of active and inactive molecules with this scaffold. As described in the methods, the *color* ranges from *red* to *blue* based on the proportion of actives. Nodes with no experimental data are *colored white*. The *color* is more vivid when there are a large number of actives within the scaffold

PME) being 0.61. The average correlation between the naive and gold standard methods is 0.57, with the highest correlation (from project TDP) being 0.83 and the lowest (from project ATA) being 0.36. The difference between the average correlations of the two methods to the gold standard indicates that the imputed method outperforms the naive method in approximating the actual presence of active scaffolds.

Correctly identified missed scaffolds

Very often, the imputed trees clearly highlight active scaffolds missed by the naive method (Fig. 6). There are several cases where none of the actives identified in the qHTS assay make it into the confirmatory test set of our simulated HTS assay, and are therefore incorrectly labeled as inactive. In these cases, the imputed number of actives at each node closely matches the actual numbers in the gold-standard trees.

This result is exciting because entirely new scaffolds are frequently identified with imputed trees. We define a

correctly predicted scaffold as one where the predicted proportion of active molecules in a scaffold is at least 50 % of the actual proportion of active molecules as reported by the gold standard method. We consider only those scaffolds that have at least one missed active molecule and at least two active molecules total. On average across all the screens, the imputed trees identify 71.7 % of missed active scaffolds. In the best case (project TDP), 90.2 % of missed active scaffolds are accurately identified, 3.9 % of the total number of active scaffolds. In the worst case (project MIR), 55.5 % of missed active scaffolds are accurately identified, 12.3 % of the total.

False negative scaffolds

Less often, the imputed trees underestimate the number of active molecules at each node (Fig. 7), and can color active molecules as if they were inactive. Normally, the imputed trees still predict that some molecules are active. Strikingly, the imputed trees are never worse than the naive trees, suggesting that, in the worst cases, imputation performs at least as well as the naive method.

Substantially under-predicting the number of actives in a scaffold is uncommon. The imputed trees under-predict the number of actives by 50 % for, on average, 28.3 % of active scaffolds that contain two or more active molecules. In the best case (project TDP), 9.8 % of scaffolds are under-predicted, which represents 0.4 % of the total active scaffolds. In the worst case (project MIR), 44.5 % of scaffolds are under-predicted, which represents 9.9 % of the total.

False positive scaffolds

Very rarely, the imputed trees color a scaffold active where the gold standard does not (Fig. 8). These cases are false positives, where the imputed tree highlights a scaffold as if were active although it is inactive. Fortunately, these cases are rare and were difficult to identify.

The imputed trees over-predict the number of actives by 10 % of the active proportion of, on average, 3.5 % of inactive scaffolds. In the best case (project HLM), 3 % of scaffolds are predicted active. In the worst case (project TDP), 12.9 % of scaffold predicted active.

Overall accuracy

Clearly, the imputed trees are only an approximation of the gold-standard trees. In prior sections, we have quantitatively assessed efficiency in identifying missed active scaffolds, false positive scaffolds, and false negative scaffolds. Another way of quantifying the overall tradeoff between these metrics is the Receiving Operator

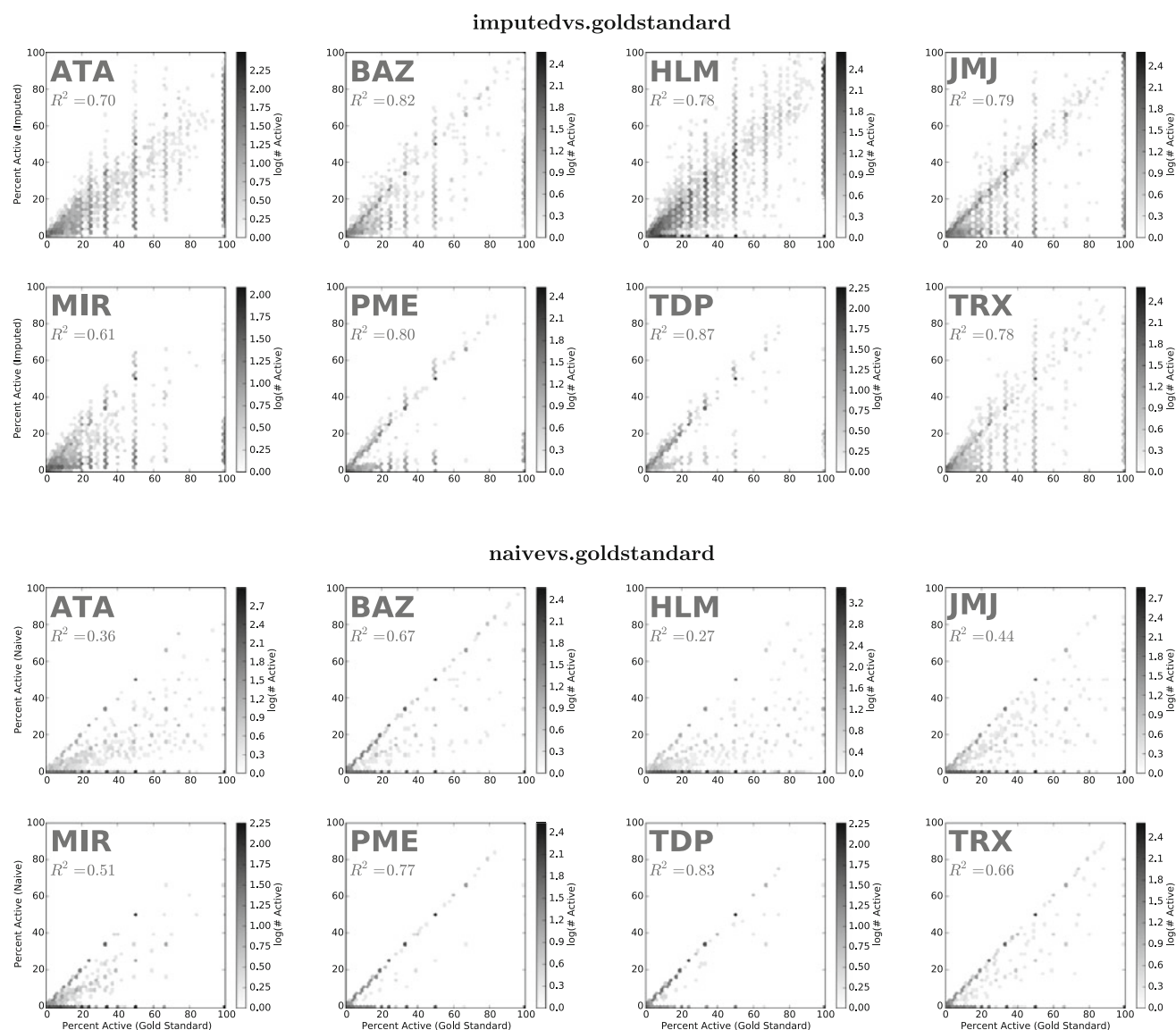


Fig. 4 Scaffold imputation accuracy. The number of actives for each scaffold as reported by the naive method are graphed against the true number of active molecules in that scaffold. The correlation is reported on each graph. The key observation is that a large number of scaffolds hug the x-axis in the naive method, representing scaffolds

where no molecules are confirmed active though they exist in the primary screen. A large proportion of these missed scaffolds shift upward, off the x-axis, in the imputed method because imputation fills in the missing measurements, correctly identifying these scaffolds as actives

Characteristic (ROC) curve (Fig. 9). Here, we order all the scaffolds in the screen by the proportion of active molecules computed using either the naive or imputed strategies. Scaffolds with one or more active molecules according to the gold-standard are considered true positives, and those with zero actives are considered true negatives.

The imputed trees dramatically outperform the naive trees. The AUC for the imputed trees is, on average, 0.97 compared with 0.73 for the naive trees. Moreover, within the first 10 % of the ROC curve, the imputed trees identify 94.6 % of the active scaffolds, compared with the 51.5 %

identified by the naive trees. In the best case (project MIR), 99.9 % of active scaffolds are identified within the first 10 % of the ROC curve, while in the worst case (project HLM), 82.2 % of active scaffolds are identified.

Discussion

Most importantly, this study demonstrates that imputed visualizations of HTS data can robustly highlight active scaffolds missed by other methods. This has important implications for drug discovery. For example,


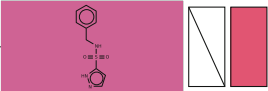
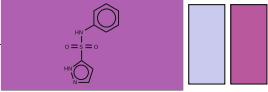
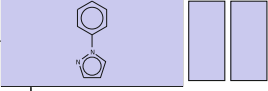

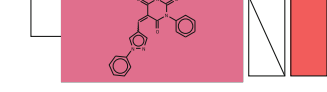
	Imputed	Naive	Gold Standard
	2.1% (3.2/153)	0.0% (0/148)	1.3% (2/153)
	69.4% (4.2/6)	(0/0)	83.3% (5/6)
	49.8% (6.0/12)	0.0% (0/4)	58.3% (7/12)
	0.5% (1.0/206)	0.0% (0/204)	0.5% (1/206)
	63.7% (8.3/13)	0.0% (0/2)	84.6% (11/13)
	79.1% (3.2/4)	(0/0)	100.0% (4/4)

Fig. 5 A representative scaffold tree. Each cell of the table displays the percentage of molecules associated with the scaffold that are active. In *parenthesis*, are the number of actives over the total number of molecules associated with the scaffold. In the naive method, these counts only reflect those molecules sent for confirmatory testing. The background hue of each node ranges from *blue* to *red* according to the proportion of active molecules in that scaffold, with a *pure blue* indicating no active molecules and a *pure red* indicating all active molecules. The *lightness* of each node is scaled by the proportion of

actives times the number of actives, so that active scaffolds are more vivid. For instance, a scaffold whose molecules are all active but that contains only three actives will be a *lighter pink* compared to a scaffold whose molecules are all active but that contains ten actives. The *color* of the molecular framework is either *black* or *white*, and is chosen to enhance visibility. To facilitate comparisons between each method, the molecule backgrounds are *shaded* with their imputed *color*, and the two swatches to the right of each node display the naive (*left*) and gold-standard (*right*) *color*

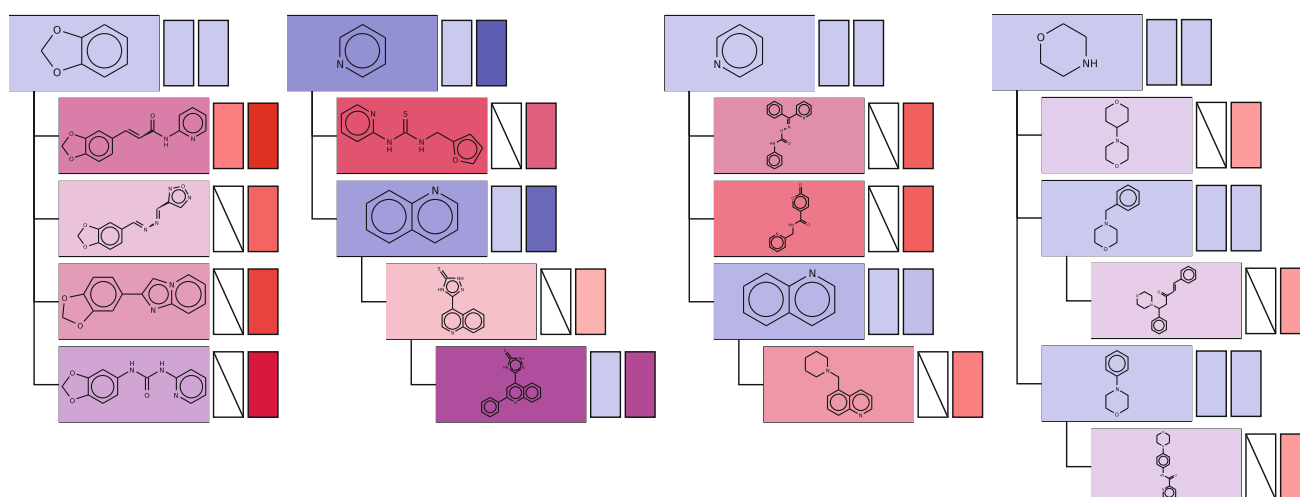


Fig. 6 Identifying missed actives. In these four examples, (taken from ATA, HLM, JMJ, and TRX from *left to right*) the naive tree fails to find highly active scaffolds—because those scaffolds are not sent for further testing—while the imputed tree clearly identifies them.

Missed active scaffolds look red in both the imputed and gold standard tree, but look blue or white in the naive tree. Trees with correctly identified missed actives are common

visualization of public screening data can be used to identify active scaffolds missed by the original screeners. These missed actives are non-obvious and can be the

starting point for generating novel intellectual property at a very low cost in comparison with the investment of running a large screen.

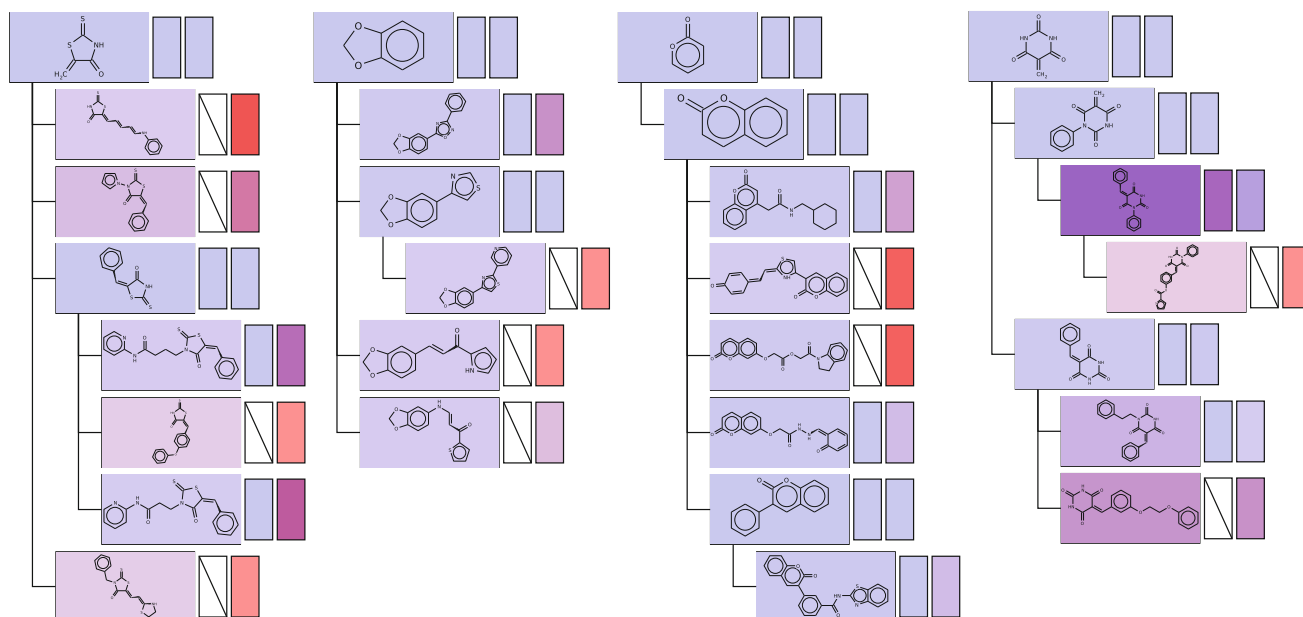


Fig. 7 False negative scaffolds. In these four examples, (taken from BAZ, MIR, PME, and TRX from *left to right*) the imputed trees are less accurate in identifying missed actives. The imputed proportion of actives is higher than the naive proportion but lower than the gold standard proportion. The *dull purples* and *blues* of the imputed

method, when compared to the reds of the gold standard, indicate that the imputed method is failing to identify missed actives in these cases. However, the naive method performs worse than the imputed method in these cases. Fortunately, false negative scaffolds like this are uncommon

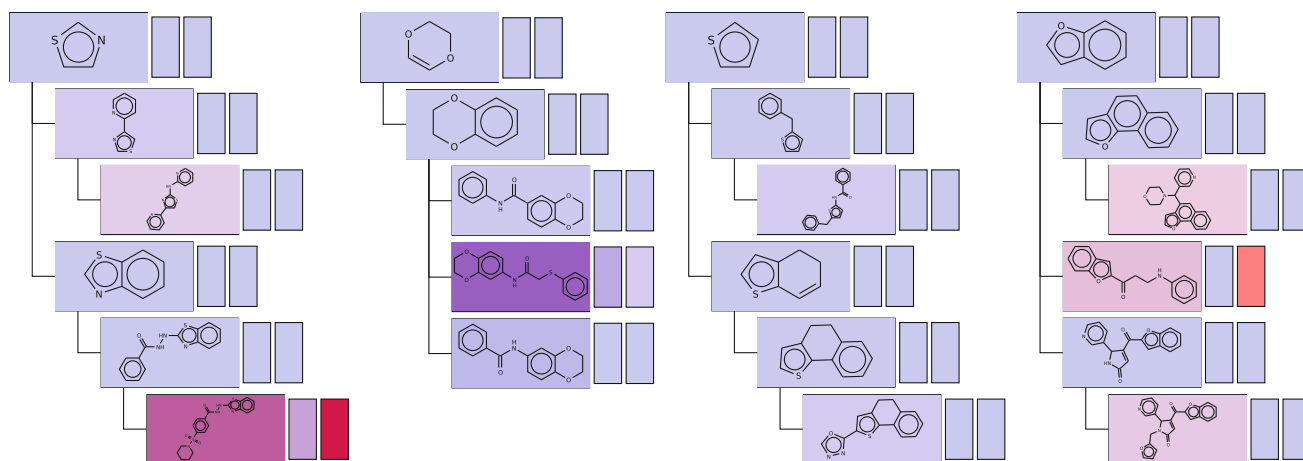


Fig. 8 False positive scaffolds. In these four examples, (taken from BAZ, MIR, TDP, and TRX from *left to right*) the imputed trees predict a substantially higher proportion of actives than the gold standard. False positive scaffolds are imputed to be marginally active when, in fact, they are not active at all. Usually, the imputed and gold-standard trees still have very similar *colors*. When the number of actives are over predicted, the difference from the real value is usually

very small. The average over-estimation across all data sets amounts to 0.155 active molecules per scaffold. This results in only a subtle difference in *colors* in most cases, as seen in the third tree from the *left*. Although the third and sixth scaffolds in this tree are predicted to have more actives than they actually do, the difference is so small that it causes an almost imperceptible change in *color*. Moreover, these cases are very rare

Although significant, these results are expected. Our finding that noisy primary screening data can yield valuable information is consistent with our prior work in structure mining [11] and related research in HTS analysis [4, 8]. For example, compound set enrichment (CSE) identifies statistically significant distributions of scaffolds in the primary screening data, enabling the detection of

groups of active molecules that were initially missed [8]. In a similar manner, local hit rate analysis (LHR) identifies clusters of molecules whose distribution is statistically significant in the primary screening data [4].

Imputation has some advantages over those methods already in the literature. First, it does not discard data from high-confidence confirmatory experiments, but cohesively

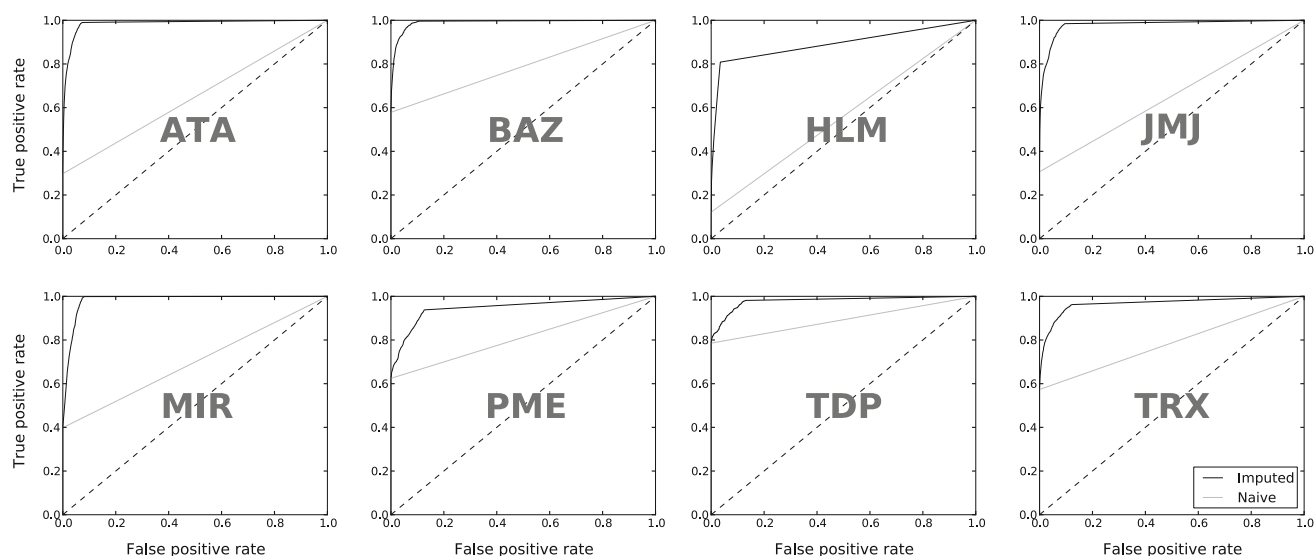


Fig. 9 ROC curves. In each project, scaffolds are ordered by the proportion of their molecules that are active. The results of the qHTS defines which scaffolds are truly active, with at least one active

integrates information from both the primary and confirmatory experiments into the same analysis. Second, by allowing for arbitrary definitions of “active” molecules, our approach may be tuned to the most appropriate definition for each project. We used a 5 μ M potency cutoff in this study, but any domain-specific definitions—including complex definitions like differential activity or passing a counter screen—are possible.

We recognize the limitations of this study. It focuses on missing dose-response measurements in a two-stage HTS experiment; however, most HTS campaigns are much more complicated, including several more stages. Of particular importance are subsequent counter-screens that filter out molecules that initially appear interesting in the dose-response stage of the campaign. Nonetheless, the same techniques used in this study can predictively model the results of subsequent counter screens. Imputation is a powerful technique that can work at all stages of the HTS experiment.

Moreover, the imputation method we use here might be improved. For brevity, we only consider NN3 in this study, but almost any predictor whose output can be interpreted as a probability may be used instead. Viable alternatives include logistic regressors, neural networks with different configurations, support-vector machines, or kernel density estimators [5, 11, 12]. We have observed that low-parameter models seem to perform better, but it is possible that performance gains could be realized by taking chemical similarity into account [31]. This direction is important and addressed in future work, but is outside the scope of this study. We intend to more thoroughly explore refinements of the imputation protocol in future work.

molecule. The imputed trees substantially outperform the naive trees in all cases, identifying the majority of active scaffolds in the early part of the curve

Finally, other formulas could be used for computing a color based on the number of active and inactive molecules associated with a scaffold. Other formulas might yield colors that more clearly highlight interesting scaffolds, but thoroughly investigating this is beyond our scope. The key point is that imputation would improve the quality of any coloring by more accurately estimating the number of active and inactive molecules associated with each scaffold. Imputation is a powerful method of managing missing data.

Conclusion

We successfully applied imputation to manage missing data in a HTS campaign. In this study, we demonstrate in several projects that imputing missed data and using an effective visualization enables the rapid identification of active scaffolds that were missed in the initial screen. Our method outperforms naive visualizations that do not impute missing data.

Acknowledgments MRB collaborated with SJS to write the initial manuscript. MRB implemented the imputed tree based on an idea by SJS and ran most of the experiments. BTC prepared the imputed data downloaded from PubChem. Edward Holson provided helpful comments and edits to the manuscript. The Pathology and Immunology Department at the Washington University in St. Louis supports BTC, MRB, and SJS. Marvin was used to generate the chemical structures in Fig. 4; Marvin 5.3.5, 2010, ChemAxon (<http://www.chemaxon.com>).

Conflict of interest The authors declare they have no conflict of interests to disclose.

References

- Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DVS, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS (2011) *Nat Rev Drug Discov* 10(3):188. <http://dx.doi.org/10.1038/nrd3368>
- Glick M, Klon A, Acklin P, Davies J (2004) *J Biomol Screen* 9(1):32. PMID: 15006146
- Glick M, Jenkins J, Nettles J, Hitchings H, Davies J (2006) *J Chem Inf Model* 46(1):193. PMID: 16426055
- Posner BA, Xi H, Mills JEJ (2009) *J Chem Inf Model* 49(10):2202–2210
- Swamidass SJ, Bittker JA, Bodycombe NE, Ryder SP, Clemons PA (2010) *J Biomol Screen* 15(6):680
- Swamidass SJ, Calhoun BT, Bittker JA, Bodycombe NE, Clemons PA (2011) *Bioinformatics* 27(16):2271–2278
- Inglese J, Auld D, Jadhav A, Johnson R, Simeonov A, Yasgar A, Zheng W, Austin C (2006) *Proc Natl Acad Sci* 103(31):11473, PMID: 16864780
- Varin T, Gubler H, Parker C, Zhang J, Raman P, Ertl P, Schuffenhauer A (2010) *J Chem Inf Model* 277–279, PMID: 21073183
- Yan S, Asatryan H, Li J, Zhou Y (2005) *J Chem Inf Model* 45(6):1784
- Lakshminarayan K, Harp S, Goldman R, Samad T, et al. (1996) *Proceedings of the second international conference on knowledge discovery and data mining*, pp 140–145
- Ranu S, Calhoun BT, Singh AK, Swamidass SJ (2011) *Mol Inf* 30(9):809. doi:[10.1002/minf.201100058](https://doi.org/10.1002/minf.201100058)
- Tanrikulu Y, Kondru R, Schneider G, So W, Bitter H (2010) *Mol Inf* 29(10):678
- Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch M, Waldmann H (2007) *J Chem Inf Model* 47(1):47
- Wang Y, Xiao J, Suzek T, Zhang J, Wang J, Bryant S (2009) *Nucleic acids research* 37 (Web Server issue), W623. PMID: 19498078
- Bolton E, Wang Y, Thiessen P, Bryant S (2008) *Annu Rep Comput Chem* 4:217. PMID: 19498078
- McCulley J, Myung K (2011) *Cell Cycle* 10:3434
- Lee KY, Yang K, Cohn MA, Sikdar N, D'Andrea AD, Myung K (2010) *J Biol Chem* 285:10362
- Jones M, Hamana N, Nezu J, Shimane M (2000) *Genomics* 63(1):40
- Quinn A, Allali-Hassani A, Vedadi M, Simeonov A (2010) *Mol Biosyst* 6(5):782
- Liu F, Chen X, Allali-Hassani A, Quinn A, Wigle TJ, Wasney GA, Dong A, Senisterra G, Chau I, Siarheyeva A et al. (2010) *J Med Chem* 53(15):5844–5857
- Lee J, Thompson J, Botuyan M, Mer G (2007) *Nat Struct Mol Biol* 15(1):109
- Sonkoly E, Wei T, Janson PC, Saaf A, Lundeberg L, Tengvall-Linder M, Norstedt G, Alenius H, Homey B, Scheynius A, Stahle M, Pivarsci A (2007) *PLoS ONE* 2:e610
- Chan JA, Krichevsky AM, Kosik KS (2005) *Cancer Res* 65:6029
- Biertumpfel C, Zhao Y, Kondo Y, Ramon-Maiques S, Gregory M, Lee JY, Masutani C, Lehmann AR, Hanaoka F, Yang W (2010) *Nature* 465:1044
- Albertella MR, Green CM, Lehmann AR, O'Connor MJ (2005) *Cancer Res* 65:9799
- Marchand C, Lea W, Jadhav A, Dexheimer T, Austin C, Inglese J, Pommier Y, Simeonov A (2009) *Mol Cancer Ther* 8(1):240
- Dexheimer T, Antony S, Marchand C, Pommier Y (2008) *Anti-cancer Agents Med Chem* 8(4):381
- Arner ES (2009) *Biochim Biophys Acta* 1790:495
- Witte AB, Anestalt K, Jerremalm E, Ehrsson H, Arner ES (2005) *Free Radic Biol Med* 39:696
- Baldi P, Brunak S (2001) *Bioinformatics: the machine learning approach*. The MIT Press, Cambridge
- Swamidass S, Azencott C, Lin T, Gramajo H, Tsai S, Baldi P (2009) *J Chem Inf Model* 49(4):756