Supplementary Note

Contents

Evaluation of feature selection algorithms

We evaluated five different commonly used feature selection algorithms on a prototype dataset (**Supplementary Methods** online). Three of the algorithms are multivariate algorithms, namely stepwise discriminant analysis[1] (SDA, **Supplementary Methods** online), genetic algorithm[2] (GA, **Supplementary Methods** online), and support vector machine recursive feature elimination[3] (**Fig. 2b** in main text). In a previous study[4], SDA and GA were found to be the best feature selection algorithms for the classification of fluorescent microscopy images. Support vector machine recursive feature elimination has been found to perform well in high dimensional data, such as mRNA microarray data[3, 5, 6]. We considered support vector machine recursive feature elimination with incremental feature elimination (SVM-RFE1) and feature subset elimination (SVM-RFE2). SVM-RFE2 removes a subset of features at a time (Profile and classification accuracy computation, **Methods** in main text) in order to decrease the computational runtime at the possible expense of feature selection accuracy. The remaining two feature selection algorithms we compared are univariate, namely the t-test (TTEST, **Supplementary Methods** online) and Kolmogorov-Smirnov test (KSTEST, **Supplementary Methods** online). Finally, the algorithm with no feature selection (NONE) was also considered. Other feature reduction algorithms, such as the principal component analysis[2] or independent component analysis[2], were not considered because they transform feature space coordinates and the resulting feature sets may not be easily interpretable.

Each of the algorithms selected a series of feature subsets with incrementally increasing cardinalities, except for NONE which selected only one feature set with all the features. A

classification accuracy score and a multivariate profile were estimated from each of these subsets using a support vector machine. The maximum and minimum classification accuracies ($C_{\max}$ and $C_{\min}$ respectively) were determined. In practice, we found that the classification accuracy associated with each profile remained relatively constant as features with little discriminatory information were removed from the whole feature set (**Fig. 2b** in main text). The trend continued until features critical for discriminating compound effects were removed, after which the classification accuracy started to decrease. We observed that the classification accuracy before this drop-off point (e.g. **Fig 2b** in main text, red box) was roughly the same as the global maximum classification accuracy value. We selected the minimum feature number with classification accuracy at least $0.9(C_{\max} - C_{\min}) + C_{\min}$, which we typically observed to be at or near this drop-off point. The difference between the accuracy at this point and the global maximum was typically around 1-3%.

Ideally, the performance of a feature selection algorithm should be assessed in term of its false positive and false negative rates in selecting relevant features. However, the identities of the relevant and irrelevant features are usually unknown in most real data. Hence the false positive and false negative feature selection rates cannot be measured directly and have to be estimated indirectly through other related criteria, such as the empirical classification accuracy of the selected features. Besides classification accuracy, we choose to add random features, which were artificially generated from noise, to the data and estimate the percentage of random features selected (number of random features selected/total number of features selected). This criterion is closely related to the feature selection false positive rate (**Supplementary Methods** online).

For this study, the performance criteria we used were: 1) classification accuracy of the selected features, 2) percentage of random features selected, 3) selected number of features and 4) computational runtime. The most desirable feature selection algorithm is the one that selects features with the highest classification accuracy at effective compound concentrations, the lowest classification accuracy at ineffective compound concentrations, and the lowest percentage of random features selected. Given feature selection algorithms that perform similarly on the above criteria, we then prefer the algorithm that produces the smallest number of features (which presumably offers the most compact and least redundant representation of the drug effect), and the lowest computational runtime.

First, we considered the classification accuracy of the selected features (**Supplementary Fig. N1**). A 3x2 cross-validation paired *t*-test (**Supplementary Methods** online) was used to test the null hypothesis that the classification accuracy of features selected by an algorithm was similar or lower than the classification accuracy of features selected by SVMRFE2. The multiple hypothesis testing was corrected by controlling the false discovery rate[7]. Algorithms selecting features with significantly higher classification accuracy were indicated (* in **Supplementary Fig. N1**, $q = 0.10$). The unadjusted *P*-values obtained from the hypothesis testing are also shown (**Supplementary Fig. N2**). The difference between average classification accuracies should not be used directly as an indication of the performance difference between two algorithms, because the average classification accuracies were estimated from the same fold of data and thus were not independent.

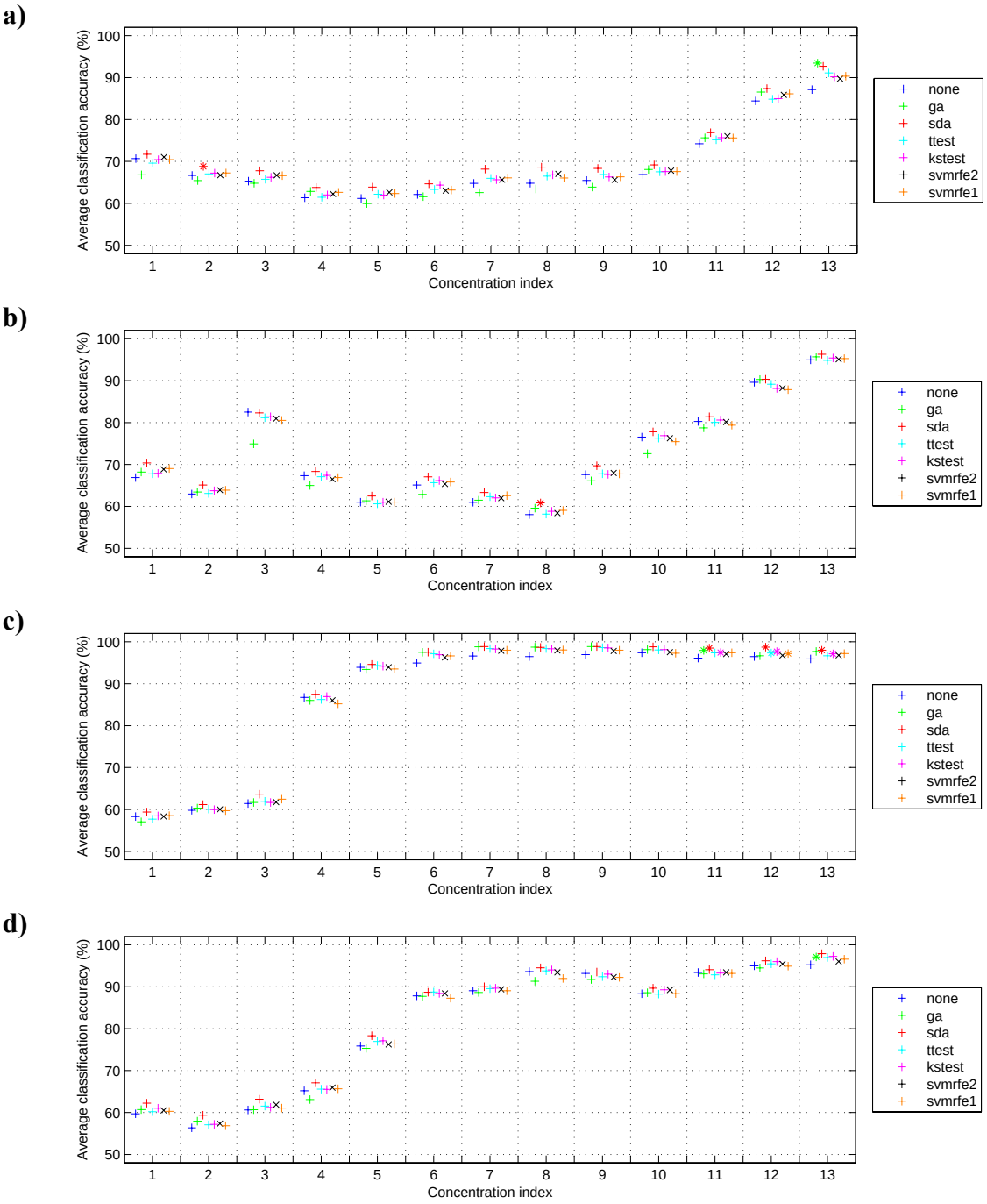
In general, the classification accuracies of the features selected by all feature selection algorithms were similar or lower than the features selected by SVMRFE2 at most of the concentrations indices (+ in **Supplementary Fig. N1, N2**). Occasionally at some concentration indices, some feature selection algorithms selected features with significantly higher classification accuracy than SVMRFE2 (* in **Supplementary Fig. N1, N2**). Most notably, the stepwise discriminant analysis (SDA) was significant at 9 concentration indices. However, 5 of the concentration indices were shown later to be having no significant compound effect (**Fig. 3** in main text, and **Supplementary Fig. N1**).

More importantly, all feature selection algorithms gave highly correlated titration curves, and should detect the drug effects at around the same concentrations. Interestingly, the algorithm with no feature selection (NONE) sometimes performed similarly as SVMRFE2. This phenomenon was observed previously for some high-dimensional datasets[6, 8], and may be attributed to the robustness of SVM classifiers in handling irrelevant features. In summary, in cases for which slightly higher classification accuracy at lower compound concentrations (usually with null drug effect) is acceptable, SDA appeared to be the winner for classification accuracy. Otherwise, all feature selection algorithms selected features with similar classification accuracies.

The differences among the algorithms became greater when considering the percentage of random features selected by these different algorithms (**Supplementary Fig. N3**). The 3x2 cross-validation paired *t*-test was used to test the null hypothesis that the percentage of random features selected by an algorithm was similar or lower than the percentage of random feature selected by SVMRFE2. On all of the prototype compounds, SDA selected significantly more random features at many concentrations (**Supplementary Fig. N3**). Furthermore, SDA selected more percentage of random features than NONE (*i.e.*, >6.76% selected features were random) on at least one of these concentrations, indicating the false positive rate of SDA could be occasionally high on our dataset. The univariate algorithms (TTEST and KSTEST) also selected significantly more percentage of random features than SVMRFE2 on all prototype compounds. On the contrary, the GA did not select any random features on the same prototype compounds. SVMRFE1/2 also did not select any random features, except for one of the prototype compounds, Taxol (**Supplementary Fig. N3c**). In order to assess if the high percentage of random features selected by SVMRFE on Taxol was a rare event, the profiles for all the 100 compounds on 4 marker sets were extracted using SVMRFE2 and the percentage of profiles with at least one random feature selected was estimated to be 0.79% ±0.41% (mean ± standard deviation) on the whole dataset. The low percentage showed that SVMRFE2 omitted the random features on almost all of the profiles (>99% on average), and thus its false positive rate may be very low.

Finally, we also evaluated the number of features selected (**Supplementary Fig. N4**) and the computational runtime (**Supplementary Fig. N5**) for all the feature selection algorithms.

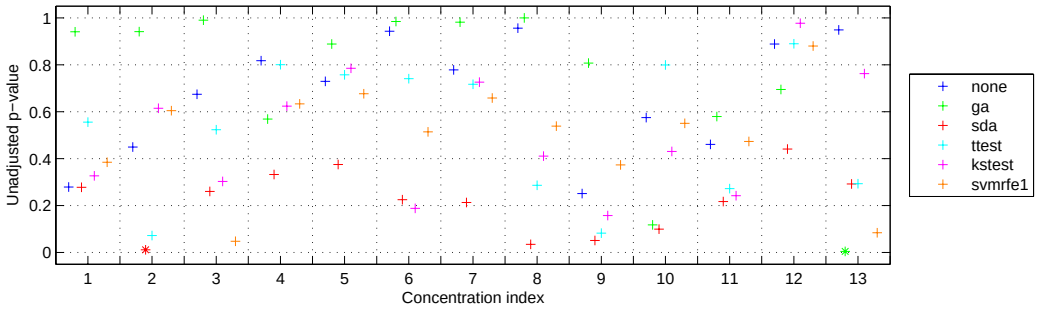
Supplementary Figure N1: Classification accuracy



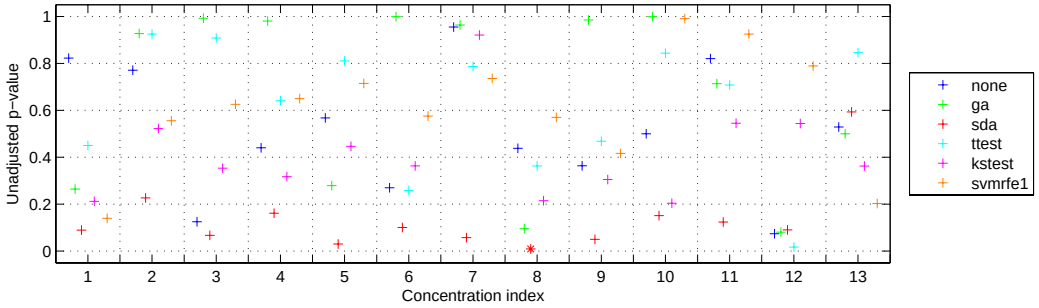
The estimated average classification accuracies of the features selected by candidate feature selection algorithms for **a)** Hydroxy Urea-2 on the DNA-cFos-p53 marker set, **b)** Oxamflatin on the DNA-pp38-pERK marker set, **c)** Taxol on the DNA-MT-actin marker set, and **d)** Campthothecan on the DNA-SC35-anillin marker set. The null hypothesis tested was the classification accuracy of features selected by an algorithm was similar or lower than the classification accuracy of features selected by SVMRFE2. (* = algorithm that rejected the null hypothesis, + = algorithm that accepted the null hypothesis, x = algorithm with constant difference to SVMRFE2, one-tailed 2x3 CV paired *t*-test, *q*-value threshold=0.10)

Supplementary Figure N2: Unadjusted *P*-values for classification accuracy comparison

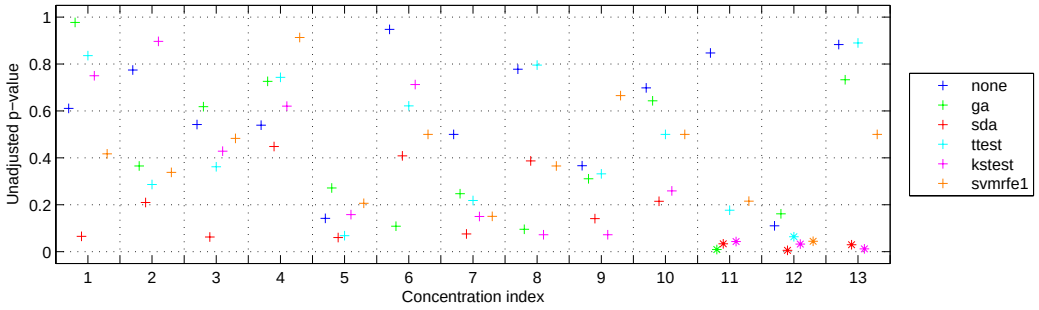
a)



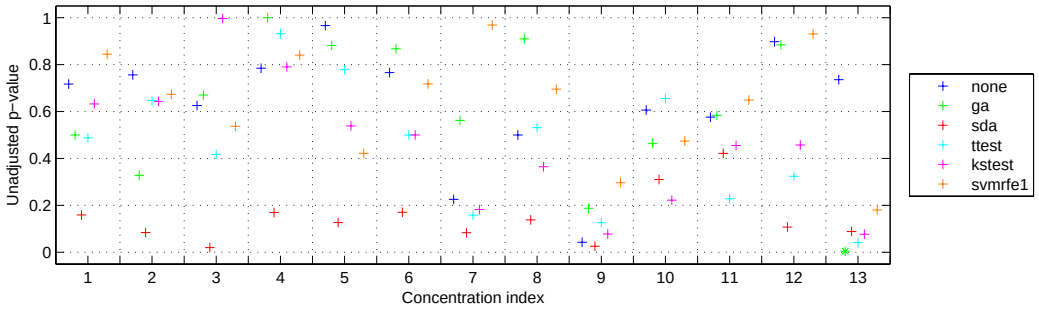
b)



c)



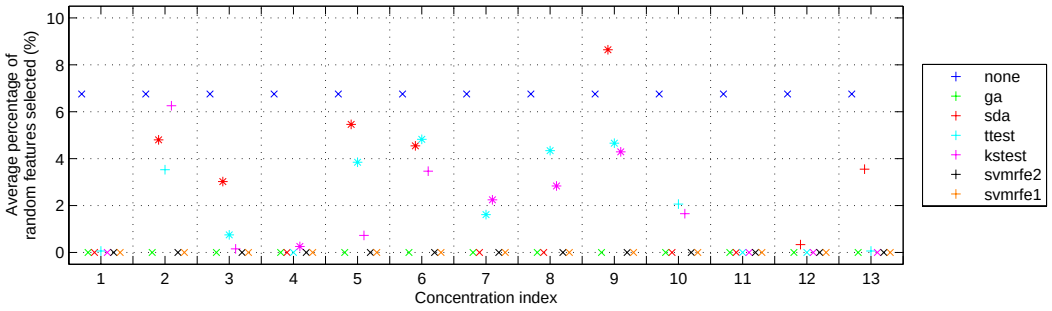
d)



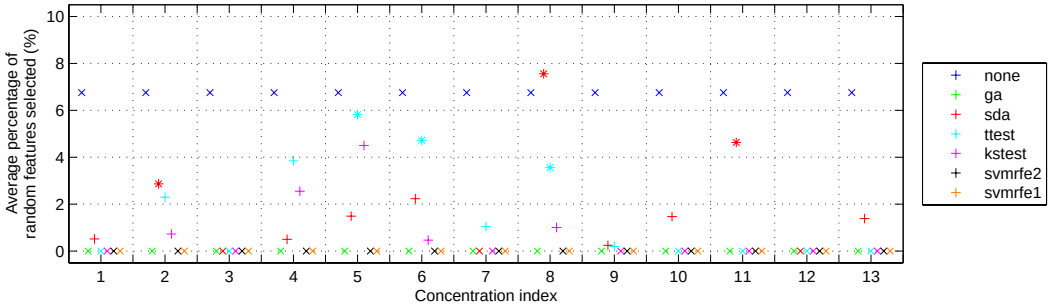
The unadjusted *P*-values for the classification accuracy comparison of candidate feature selection algorithms to SVMRFE2 on **a)** Hydroxy Urea-2 on the DNA-cFos-p53 marker set, **b)** Oxamflatin on the DNA-pp38-pERK marker set, **c)** Taxol on the DNA-MT-actin marker set, and **d)** Campthothecan on the DNA-SC35-anillin marker set. The null hypothesis tested was the classification accuracy of features selected by an algorithm was similar or lower than the classification accuracy of features selected by SVMRFE2. (* = algorithm that rejected the null hypothesis, + = algorithm that accepted the null hypothesis, x = algorithm with constant difference to SVMRFE2, one-tailed 2x3 CV paired *t*-test, *q*-value threshold=0.10)

Supplementary Figure N3: Percentage of random features selected

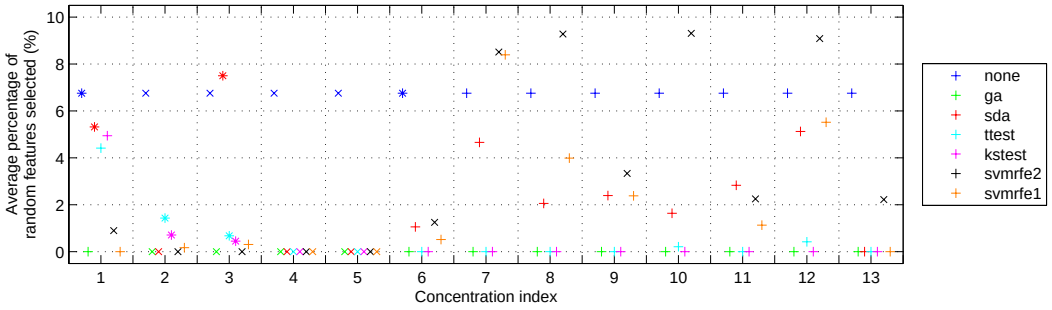
a)



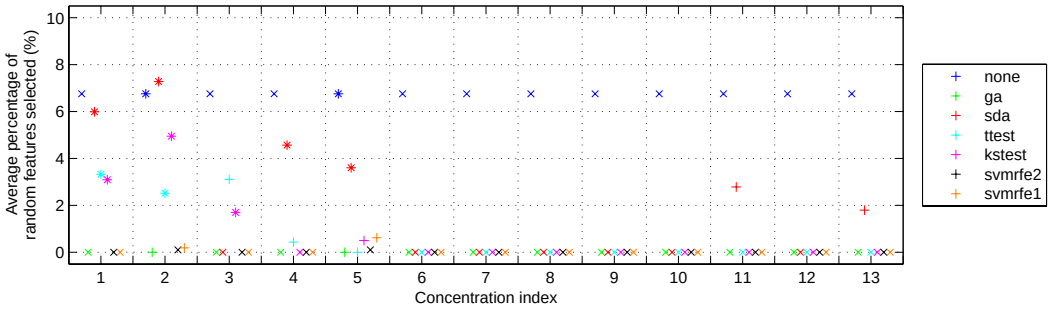
b)



c)



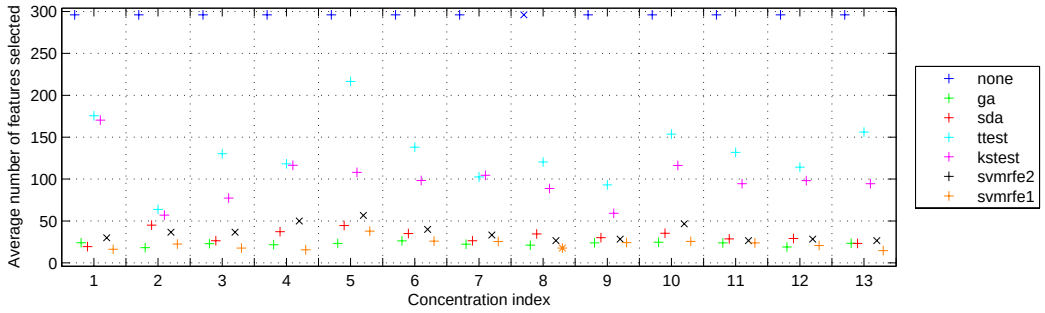
d)



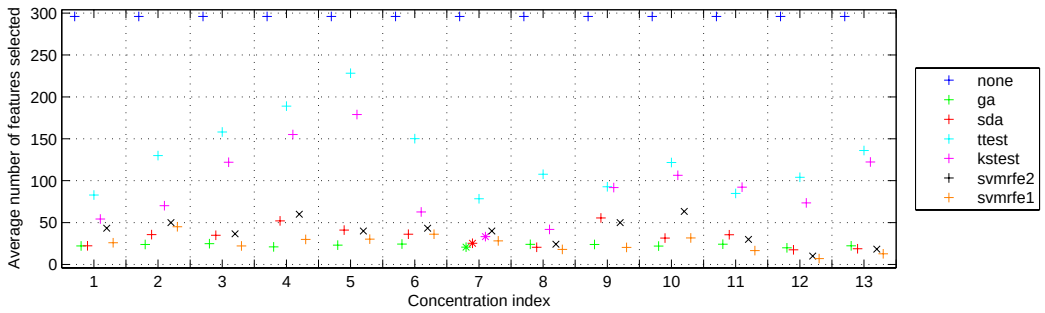
The estimated average percentages of random features selected by candidate feature selection algorithms for **a)** Hydroxy Urea-2 on the DNA-cFos-p53 marker set, **b)** Oxamflatin on the DNA-pp38-pERK marker set, **c)** Taxol on the DNA-MT-actin marker set, and **d)** Campthothecan on the DNA-SC35-anillin marker set. The null hypothesis tested was the percentage of random features selected by an algorithm was similar or lower than the percentage of random features selected by SVMRFE2. (* = algorithm that rejected the null hypothesis, + = algorithm that accepted the null hypothesis, x = algorithm with constant difference to SVMRFE2, one-tailed 2x3 CV paired *t*-test, *q*-value threshold=0.10)

Supplementary Figure N4: Number of features selected

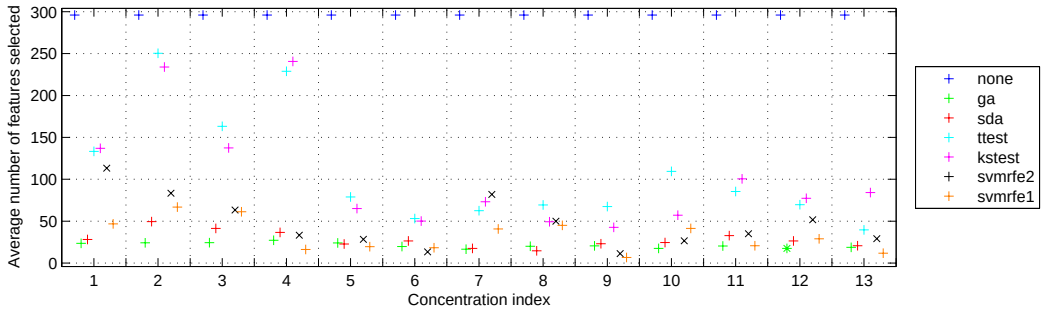
a)



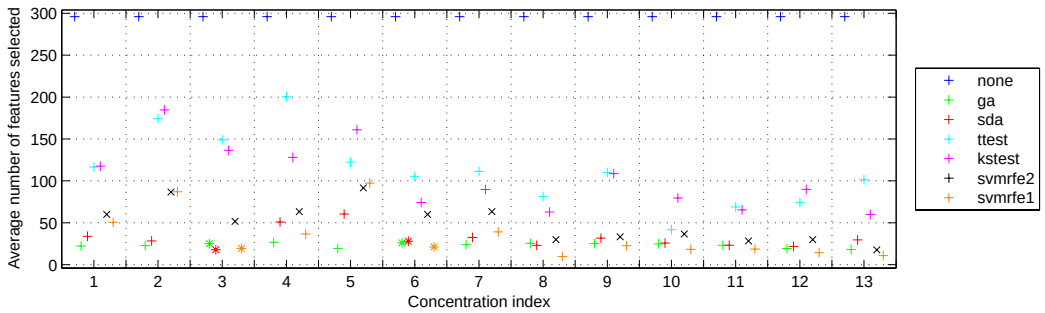
b)



c)

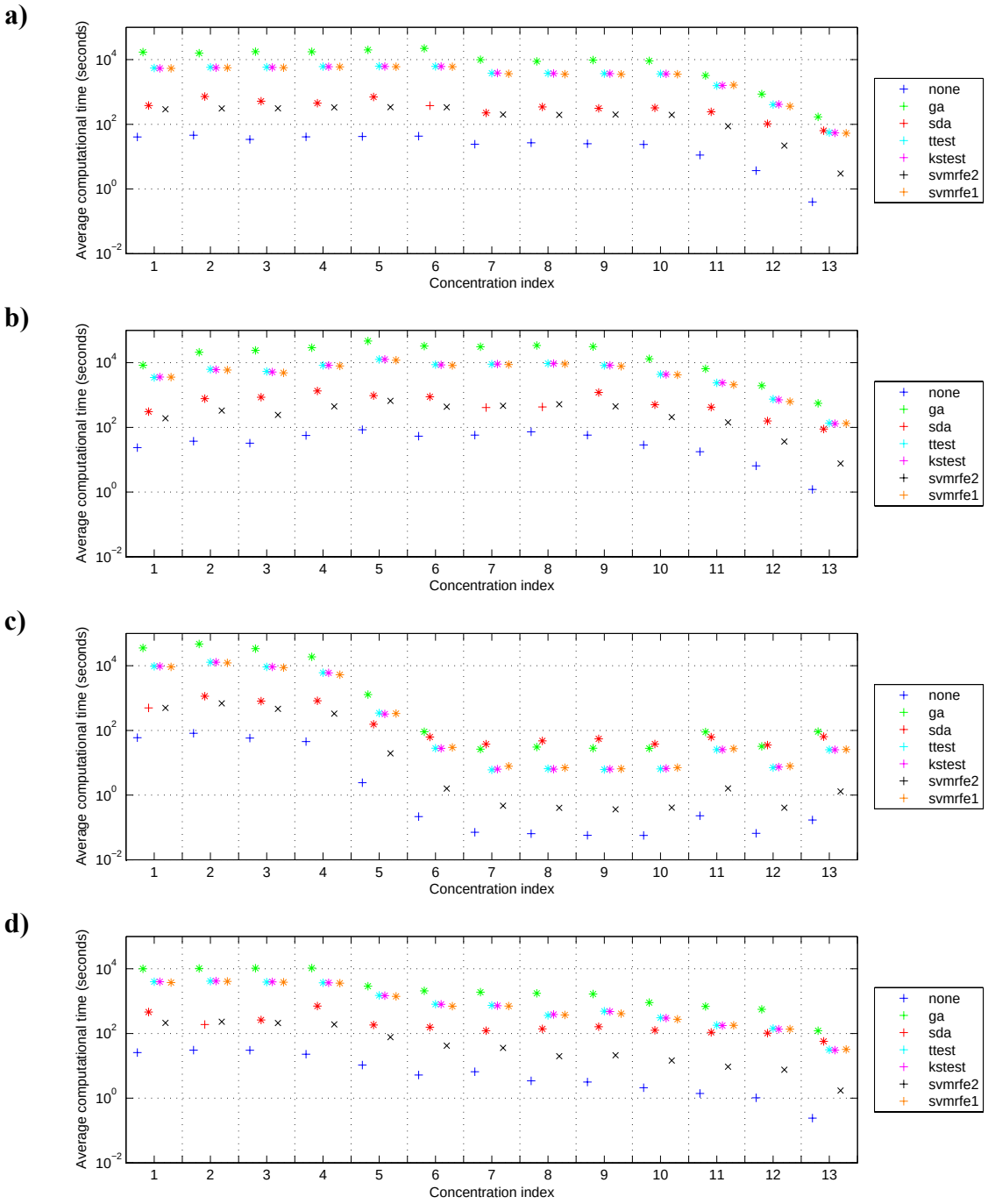


d)



The estimated average number of features selected by by candidate feature selection algorithms for **a)** Hydroxy Urea-2 on the DNA-cFos-p53 marker set, **b)** Oxamflatin on the DNA-pp38-pERK marker set, **c)** Taxol on the DNA-MT-actin marker set, and **d)** Campthothecan on the DNA-SC35-anillin marker set. The null hypothesis tested was the number of features selected by an algorithm was similar or higher than the number of features selected by SVMRFE2. (* = algorithm that rejected the null hypothesis, + = algorithm that accepted the null hypothesis, x = algorithm with constant difference to SVMRFE2, one-tailed 2x3 CV paired *t*-test, *q*-value threshold=0.10)

Supplementary Figure N5: Computational runtime



The estimated average computational runtime of candidate feature selection algorithms for **a)** Hydroxy Urea-2 on the DNA-cFos-p53 marker set, **b)** Oxamflatin on the DNA-pp38-pERK marker set, **c)** Taxol on the DNA-MT-actin marker set, and **d)** Campthothecan on the DNA-SC35-anillin marker set. The null hypothesis tested was the computational runtime of an algorithm was similar or lower than the computational runtime of SVMRFE2. (* = algorithm that rejected the null hypothesis, + = algorithm that accepted the null hypothesis, x = algorithm with constant difference to SVMRFE2, one-tailed 2x3 CV paired *t*-test, *q*-value threshold=0.10)

Evaluation of clustering validation algorithms

We compared four clustering validation algorithms based on silhouette[9], gap statistics[10], resampling[11, 12], and perturbation (**Supplementary Methods** online). Clustering validation algorithms based on AIC/BIC was not considered because the estimation of maximum likelihood could not be performed accurately for small number of data (~3-4 profiles/cluster for a compound). A previous study[10] showed that gap statistics gave good results on simulated normally-distributed data with dimension higher than the number of samples, but it is unclear how gap statistics would perform on small number of profiles.

Since the actual number of clusters and the generation model of the data were unknown, a human expert's decision was used as the "gold standard". The decision was made based on examination of the calculated pairwise dissimilarity scores and the multi-dimensional scaling plot of the profiles. When the separation between clusters was not obvious or there was any doubt about the number of clusters, a "null decision" was made. For the comparison, the number of clusters was restricted to be from 2 to 5, except for the gap statistics, where the minimum number of cluster is 1.

The result for the comparison is shown in **Supplementary Table N1**. Taxol was omitted because the human expert could not make a decision. On the other three prototypes, the number of clusters selected by the perturbation-based algorithm closely matched the human expert decision, followed by the gap statistics and the silhouette. Thus the perturbation-based clustering validation algorithm was chosen.

Supplementary Table N1: Evaluation of clustering validation algorithms

Prototype	Silhouette	Gap-statistics	Resampling	Perturbation	Human Expert
Camptothecin	4	**3**	5	**3**	**3**
Taxol	2	3	5	2	*
Oxamflatin	5	1	5	**3**	**3**
Hydroxy Urea-2	**2**	**2**	3	**2**	**2**

* = Number of clusters undecided, **Bold** = correct number of clusters

Comparison between univariate and multivariate approaches

We also compared the drug screening performance of our multivariate approach to a previous univariate approach based on Kolmogorov-Smirnov statistics and a Titration Invariant Similarity Score (TISS)[13]. Compound comparisons within and between compound categories[13] showed that our multivariate approach extracted more discriminative profiles than the univariate approach (**Supplementary Table N2**). For example, the univariate approach, using the combination of 4 marker sets, was unable to distinguish vesicle trafficking inhibitors from other compound categories (P = 0.206) while the new multivariate approach distinguishes this category from other categories on each of the marker sets (P = 0.001, P = 0.008, and P = 0.005 for DNA-Anillin-SC35, DNA-p53-cFos, and DNA-MT-actin respectively). The multivariate approach additionally detected five other compound categories that were completely missed by the univariate approach (calcium regulation, MAPK/p38 pathway, PI3K pathway, PKC, and protein degradation inhibitors).

Supplementary Table N2: Performance comparison between univariate and multivariate approaches

	Multivariate d-Profiles				Univariate TISS
Compound category	**DNA-Anillin-SC35**	**DNA-p53-cFos**	**DNA-pp38-pERK**	**DNA-MT-Actin**	**All marker sets**
Actin	0.001 *	0.009 *	0.405	0.020 *	0.025 *
Calcium regulation	0.043 *	0.038 *	-	-	-
Cyclooxygenase	-	0.063	0.466	0.219	-
DNA replication	0.002 *	0.001 *	0.002 *	0.004 *	0.011 *
Histone deacetylase	<0.001 *	0.004 *	0.002 *	<0.001 *	0.001 *
Kinase	0.019 *	0.028 *	0.141	0.181	0.223
Kinase; CDK	0.049 *	0.071	0.237	0.012 *	0.057
Kinase; MAPK/p38 pathway	-	0.008 *	-	-	-
Kinase; PI3K pathway	-	0.009 *	0.002 *	0.063	-
Kinase; PKC	0.002 *	0.015 *	-	0.046 *	-
Microtubule	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *
Neurotransmitter	-	-	0.202	0.137	-
Nuclear receptor	0.077	0.224	-	0.440	-
Protein degradation	0.003 *	0.139	-	0.104	-
Protein synthesis	<0.001 *	0.001 *	0.007 *	0.066	<0.001 *
Topoisomerase	0.005 *	0.298	0.430	0.475	0.005 *
Vesicle trafficking	0.001 *	0.008 *	0.114	0.005 *	0.206
Significant categories	12	11	5	7	6

Our multivariate approach was compared to a previous univariate approach based on Kolmogorov-Smirnov statistics and a Titration Invariant Similarity Score (TISS)[13]. The P-values of the comparisons within and between compound categories using the rank sum test[13] were shown. (* = Category with statistically significant rank sum values (P < 0.05), - = category with no selected d-profile or TISS score).

Drug screening performance of selected profiles

For each marker set and compound category, we selected 2-3 representative on-target d-profiles with maximum average drug screening performance (**Methods** in main text). The exclusion of off-target effects enabled the selection of on-target d-profiles from five compound categories not found significant in the drug screening process described in **Table 1** in main text (**Supplementary Table N3**).

Supplementary Table N3: Drug screening performance of selected profiles

	DNA-Anillin-SC35			DNA-p53-cFos			DNA-pp38-pERK			DNA-MT-Actin			Significant
Compound category	***n***	**AUC**	***P***	***n***	**AUC**	***P***	***n***	**AUC**	***P***	***n***	**AUC**	***P***	**Marker sets**
Actin	3	0.986	(<0.01) *	3	0.953	(<0.01) *	3	0.887	(0.01) *	3	0.981	(0.01) *	4
Calcium regulation	2	0.886	(0.03) *		-			-			-		1
Cholesterol	2	1.000	(0.04) *	2	1.000	(0.04) *		-		2	1.000	(0.04) *	3
Cyclooxygenase		-		3	0.913	(0.01) *		-			-		1
DNA replication	3	0.997	(0.01) *	3	0.996	(0.01) *	3	0.997	(0.01) *	3	0.989	(0.01) *	4
Energy metabolism	2	1.000	(0.04) *		-			-			-		1
Histone deacetylase	3	1.000	(<0.01) *	3	0.865	(0.02) *	3	0.993	(<0.01) *	3	0.999	(<0.01) *	4
Kinase		-			-			-		3	0.892	(0.01) *	1
Kinase; CDK	3	0.986	(0.01) *	2	0.929	(0.02) *	3	0.967	(0.01) *	2	0.811	(0.03) *	4
Kinase; MAPK/ERK pathway		-			-		2	1.000	(0.04) *		-		1
Kinase; MAPK/p38 pathway		-		3	0.904	(0.03) *		-			-		1
Kinase; PI3K pathway		-			-		3	0.960	(0.01) *		-		1
Kinase; PKA		-			-			-			-		0
Kinase; PKC		-			-			-			-		0
Microtubule	3	0.940	(<0.01) *	3	0.955	(<0.01) *	3	0.903	(<0.01) *	3	0.974	(<0.01) *	4
Neurotransmitter		-			-		3	0.997	(0.01) *	2	0.912	(0.02) *	2
Nuclear receptor		-			-			-			-		0
Protein degradation	3	0.930	(0.02) *		-			-			-		1
Protein synthesis	3	0.992	(<0.01) *	3	0.947	(<0.01) *	3	0.957	(<0.01) *	3	0.971	(<0.01) *	4
RNA		-			-			-			-		0
Topoisomerase	3	0.978	(0.01) *		-		2	0.894	(0.01) *		-		2
Vesicle trafficking	3	0.885	(0.01) *	3	0.958	(<0.01) *		-		3	0.960	(0.01) *	3
Significant categories			12			10			10			10	

For each category, the average drug screening performance was recomputed from 2 to 3 representative on-target d-profiles (Methods). (* = Category with statistically significant average AUC values ($P < 0.05$), - = category with no selected d-profile).

References

1. Jenrich, R.I. Stepwise Discriminant Analysis. *Statistical Methods for Digital Computers*, 76-95 (1977).
2. Duda, R.O., Hart, P.E. & Stork, D.G. Pattern Classification, Edn. 2nd. (John Wiley & Sons, New York; 2001).
3. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389-422 (2002).
4. Huang, K., Velliste, M. & Murphy, R.F. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proceedings of SPIE* **4962**, 307-318 (2003).
5. Rakotomamonjy, A., Guyon, I. & Elisseeff, A. Variable Selection Using SVM-based Criteria. *Journal of Machine Learning Research* **3**, 1357-1370 (2003).
6. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS* **98**, 15149-15154 (2001).
7. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
8. Chen, Y.-W. & Lin, C.-J. in Feature Extraction: Foundations and Applications. (eds. I. Guyon, S. Gunn, M. Nikravesh & L.A. Zadeh) 315-323 (Springer, 2006).
9. Kaufman, L. & Rousseeuw, P.J. Finding groups in data: An introduction to Cluster Analysis. (Wiley, New York; 1990).
10. Tibshirani, R., Walther, G. & Hastie, T. Estimating the Number of Clusters in a Dataset Via the Gap Statistic. *Journal of the Royal Statistical Society B* **63**, 411-423 (2001).
11. Lange, T., Roth, V., Braun, M.L. & Buhmann, J.M. Stability-Based Validation of Clustering Solutions. *Neural Comp.* **16**, 1299-1323 (2004).
12. Roth, V., Lange, T., Braun, M. & Buhmann, J. A Resampling Approach to Cluster Validation. *COMPSTAT 2002 - Proceedings in Computational Statistics*, 123-128 (2002).
13. Perlman, Z.E. et al. Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194-1198 (2004).