Journal of
Chemical Information
and Modeling

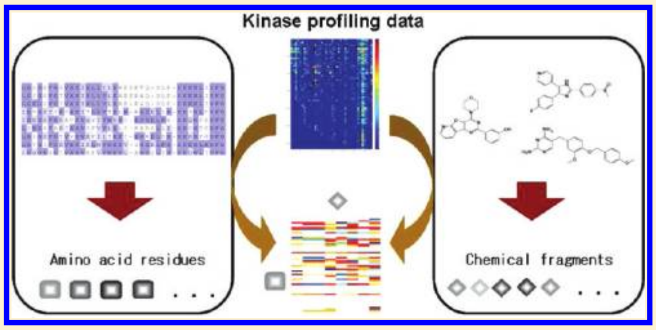
Article

pubs.acs.org/jcim

Dissecting Kinase Profiling Data to Predict Activity and Understand Cross-Reactivity of Kinase Inhibitors

Satoshi Niijima,* Akira Shiraishi, and Yasushi Okuno*

Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan



ABSTRACT: The development of selective and multitargeted kinase inhibitors has received much attention, because cross-reactivity with unintended targets may cause toxic side effects, while it can also give rise to efficacious multitargeted drugs. Here we describe a deconvolution approach to dissecting kinase profiling data in order to gain knowledge about cross-reactivity of inhibitors from large-scale profiling data. This approach not only enables activity predictions of given compounds on a kinome-wide scale, but also allows to extract residue−fragment pairs that are associated with activity. We demonstrate its effectiveness using a large-scale public chemogenomics data set and also apply our proposed model to a recently published bioactivity data set. We further illustrate that the preference of given compounds for kinases of interest is better understood by residue−fragment pairs, which could provide both biological and chemical insights into cross-reactivity.

INTRODUCTION

The human kinome, the protein kinase family in the human genome, consists of 518 genes, constituting one of the largest protein families.[1−3] Protein kinase-mediated signaling pathways are implicated in a variety of diseases such as cancer, inflammation, and diabetes.[4−6] Thus, numerous protein kinases have emerged as potential therapeutic targets, and several small molecule kinase inhibitors are currently in clinical use.[5]

Cross-reactivity with unintended targets may cause toxic side effects, while it can also give rise to efficacious multitargeted drugs. Therefore, the design of inhibitors modulating the activity of intended targets, referred to as “targeted polypharmacology”, has attracted increasing attention.[6,7] Despite intense research, however, the development of selective and multitargeted kinase inhibitors remains challenging due to the fact that the binding sites of most kinases, particularly the ATP-binding pocket which is targeted by the vast majority of existing kinase inhibitors, are highly conserved in sequence and structure.[8]

Recent advances in high-throughput screening technologies have enabled bioactivity profiling of hundreds of compounds against a panel of protein kinases,[9−16] and the so-called kinase panel has emerged as a promising approach to address the cross-reactivity issue.[17] Obviously, the kinase panel facilitates not only the identification of previously unknown targets of the compounds but also the assessment of polypharmacological profiles.[18] However, the acquisition of a kinase panel for a large number of compounds is still costly and time-consuming, and hence, the coverage of chemical space is critically limited. This could hinder the discovery of potent compounds with better selectivity and polypharmacological profiles.

In silico modeling approaches to predict activity for large-scale compound libraries on a kinomewide scale provide a powerful tool for cost-efficient virtual profiling, thereby enabling the extrapolation and augmentation of experimentally measured profiling data. In particular, a chemogenomics approach based on statistical machine learning that leverages kinase profiling data holds great promise in the optimization of selectivity and cross-reactivity.[19−24] While previous statistical models have primarily focused on the prediction of selectivity and cross-reactivity on a kinomewide scale,[20−23] it is of greater importance to extract knowledge from kinase profiling data and transfer the knowledge into rationally designing new selective and multitargeted inhibitors against intended kinases. Nevertheless, the question of how to gain knowledge about cross-reactivity from large-scale profiling data remains underexplored.

As revealed by the study of Karaman et al.,[12] the same compound does not necessarily inhibit closely related kinases, but does inhibit distantly related ones. On one hand, compound specificity to kinases is often attributable to single residue differences.[19] Indeed, a single residue mutation causes lack of specificity, and this can lead to drug resistance.[25] This fact corroborates the need to encode kinases at the amino acid residue level, which affords the key to explaining selectivity. On the other hand, compound specificity varies widely among inhibitors. For example, two different compounds that contain quinazoline as a common chemical scaffold represent both highly selective (e.g., Lapatinib) and cross-reactive (e.g., Erlotinib) inhibitors.[18] This variation cannot be attributed only to general scaffolds, suggesting the need for more elaborate encoding of compounds at the chemical fragment level.

To meet both needs, we developed a deconvolution approach to dissecting kinase profiling data with the aim of

Received: December 17, 2011
Published: March 13, 2012

better interpreting and capturing cross-reactivity. Earlier work on the chemogenomics approach to kinase selectivity focused on single residue differences to identify selectivity-determining residues for individual inhibitors,[19] while another related work focused on different fragment compositions to identify selectivity-determining chemotypes across a select set of kinases.[21] Sheridan et al.[22] aimed to predict the overall similarity of kinase pairs in terms of binding profiles and, hence, does not allow predictions on selectivity of any given compound. Unlike these existing approaches, we deconstructed kinases into amino acid residues and compounds into chemical fragments, thereby representing kinase−inhibitor pairs by a set of residue−fragment pairs. Our proposed approach not only enables activity predictions of given compounds on a kinomewide scale but also allows extraction of residue−fragment pairs that are associated with activity. In particular, we construct a dual-component naive Bayes (DCNB) model and dual-component support vector machines (DCSVMs) and demonstrate that these models are effective for activity prediction using a large-scale public chemogenomics data set. We also apply DCSVMs to a recently published bioactivity data set for external validation and further illustrate that the preference of given compounds for kinases of interest is better understood by residue−fragment pairs.

■ METHODS

Data Sets. We extracted kinase bioactivity data from the Kinase SARfari database version 3.00,[26] which is arguably the largest public knowledgebase on chemogenomics for protein kinases. This database represents a rich resource because it provides curated data sets comprising ChEMBL SAR data from the literature. The data set used in this study encompasses 342 human kinase domains, 26 627 compounds, and 85 908 bioactivity data points (e.g., IC_{50}, K_i, and K_d values), each measuring the binding affinity of a compound against a target kinase. Note that, among the possible kinase−inhibitor pairs, the experimental values were available only partially, because this database is a compilation from diverse literature sources.

This data set was first used for the internal validation of machine learning models. For this purpose, we dichotomized kinase−inhibitor pairs into two classes (actives and inactives) on the basis of the values of binding affinity. In particular, two thresholds were set for the dichotomization: <1 μM and ≤10 μM for actives and otherwise presumed inactives. The submicromolar value (<1 μM) is often used as a stringent criterion for affinity, whereas ≤10 μM offers an empirical criterion for detecting hit compounds. In the present study, we explore the two thresholds to evaluate how these would affect the prediction performance.

To further validate our approach, we used the kinomics screening data of Metz et al.,[7] which provides invaluable data for external validation, because it was very recently publicized and collected independently of the SARfari data set. This screening data originally contains >150 000 kinase inhibitory values, consisting of >3800 compounds tested against 172 different protein kinases. Of these, the inhibitory values for 1497 compounds against the 172 kinases are publicly available. We used a cutoff of <1 μM as suggested in the polypharmacological study by Metz et al.[7]

Encoding Kinases. First, we deconstructed the amino acid sequences of kinases into residues. As previously mentioned, the ATP-binding site is the primary target of the majority of existing kinase inhibitors. In accordance with this, previous studies[22,27] have focused their analysis on the residues in and around the ATP-binding region. In our analysis, we exploit information about the ATP binding site residues for most human kinases previously identified by structure and sequence-based approaches,[27] based on the premise that those residues should play a relevant role in inhibitor binding. Specifically, we used here the alignments of the ATP-binding site, consisting of 36 residues, for 469 kinases as provided by Huang et al.[27] Of the 342 kinase domain sequences, we used 341 alignable with at least one of the 469 kinases.

The amino acid residues were encoded by two methods. The first is to encode the residues directly: for each position of the ATP-binding site, the residues were encoded by fingerprints representing the presence or absence of the 20 amino acids. The second method is to encode the residues on the basis of the physicochemical properties thereof. Here we exploited the categorization of residues by the PROFEAT server.[28] The properties used were hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure, and solvent accessibility. The residues of the ATP-binding site were encoded according to three groups for each property (see Li et al.[28]). These methods are henceforth referred to as “direct encoding” and “property encoding”.

Encoding Compounds. Next, we deconstructed compounds into chemical fragments. Specifically, we fragmented each compound by the RECAP (Retrosynthetic Combinatorial Analysis Procedure) algorithm[29] with some extensions.[30] We employed the RECAP algorithm as implemented by RDKit.[31] Importantly, unlike the original RECAP algorithm, this implementation applies the fragmentation rules in a hierarchical and exhaustive manner.[30] Of the 26 627 compounds, we eliminated those that could not be fragmented and large compounds consisting of >200 fragments, resulting in 25 018 compounds. The RECAP fragmentation of these compounds generated a total of 15 604 fragments (occurring in more than one compound). We also used extended connectivity fingerprints (ECFP_6) (as calculated by Pipeline Pilot[32])[33] for the fragmentation. The number of fragments generated from the 26 627 compounds was 56 837 (occurring in more than one compound). While the ECFP_6 fragments allow elaborate description of chemical structures, the RECAP fragments may be more useful in terms of chemical synthesis and easier to interpret.

Pairwise Fragments. To better interpret and capture kinase cross-reactivity, we represent kinase−inhibitor pairs by a set of residue−fragment pairs. We refer to a combined fingerprints consisting of amino acids residues and chemical fragments as “pairwise fragments”. The fingerprinting of binding sites has been previously proposed.[34−36] However, this approach is based on the binding site information obtained from available complex structures and also does not incorporate information about chemical structures. Chen et al.[37] recently proposed a server called SiMMap for statistically deriving site-moiety maps which contain information about both compounds and proteins, but SiMMap also relies on complex structures. By contrast, our pairwise fragments do not require three-dimensional crystallographic structures and are obtained from kinase profiling data, thus substantially expanding the coverage of both the kinome space and chemical space.

Machine Learning Models for Activity Prediction. *Dual-Component Naive Bayes Model.* Bayesian modeling has been receiving much attention in virtual screening[38,39] and particularly, naive Bayes (NB) models using fragments or fragment fingerprints have proven useful with its simplicity and ease of interpretability.[40,41] Notably, NB models have shown tolerance to noisy data, and become one of the leading machine learning techniques for ligand-based virtual screening.[38]

It should be noted that previous ligand-based NB models have treated each target independently. The multiple-category modeling approach proposed by Nidhi et al.[41] also constructs a single NB model for each target, enabling simultaneous predictions against which targets a given compound exhibits activity. Unlike the ligand-based NB models, our proposed dual-component naive Bayes (DCNB) models couple amino acid residue and chemical fragment, thereby extending the NB models based on chemical fragments alone to accommodate multiple targets.

In the following, we briefly describe the DCNB model, which is based on the Laplacian-corrected NB model.[40] Given a residue r of kinase and a fragment f of compound in active kinase−inhibitor pairs, the Laplacian-corrected estimator of the posterior probability is given as

$$\Pr(\text{active}|(r, f)) = \frac{A + \Pr(\text{active})K}{A + I + K}$$

where A and I are the frequency of a residue−fragment pair (r, f) occurring in active and inactive pairs, respectively, and Pr(active) can be simply estimated as the ratio between the number of active pairs and the total number of pairs. $K = 1/\Pr(\text{active})$ is a constant of the Laplacian correction added in order to stabilize the estimator. The final estimator of the posterior probability for a kinase−inhibitor pair (p,c) can be computed as

$$\Pr(\text{active}|(p, c)) = \prod_{(r,f)} \frac{\Pr(\text{active}|(r, f))}{\Pr(\text{active})}$$

where (r, f) runs over the residue−fragment pairs occurring in (p, c). To predict whether (p, c) is active or inactive, we may simply compare $\Pr(\text{active}|(p, c))$ and $\Pr(\text{inactive}|(p, c))$, the latter being the estimator derived from inactive kinase−inhibitor pairs.

The DCNB model has the same advantages as the NB model does. A major difference may be the number of features to be handled, since residues form pairs with each chemical fragment. Still, handling a very large number of features (even more than millions) is feasible for the DCNB model.

Dual-Component SVM Model. The major limitation of the DCNB model is the assumption that the features are mutually independent, which usually does not hold. If the composite effect of features is crucial for binding, prediction performance can be improved by allowing for it. Along this line, we explore the applicability of support vector machines (SVMs),[42] a state-of-the-art machine learning method which can allow for the mutual dependence of features.

In this study, we extend ligand-based SVMs to dual-component SVMs (DCSVMs). The concept is the same as SVM models based on target-ligand kernels,[43−45] but our contribution consists in how to handle the large number of pairwise fragments. Importantly, we show that the Tanimoto kernel[46] of the pairwise fragment fingerprints can be efficiently computed by using the kernel trick.[47]

SVMs aim to find a decision function that well separates training samples from different class labels by first projecting the samples into a feature space and then building a hyperplane in that space. The decision function can be used to make predictions for test samples with unknown labels. Both training and test samples can be efficiently projected into the feature space via kernel functions,[47] which play a pivotal role in the SVM training.

DCSVMs refer to SVM models with a target-ligand kernel devised for the above-mentioned pairwise fragments. Here we use the Tanimoto kernel[46] of the pairwise fragments as the target-ligand kernel. The commonly used Tanimoto kernel function for chemical fingerprints alone can be defined as

$$k(c, c') = \frac{|c \cap c'|}{|c \cap c| + |c' \cap c'| - |c \cap c'|}$$

where $c \cap c'$ indicates the number of fingerprints that compounds c and c' have in common.

In the case of the pairwise fragments for both kinases and compounds, we have

$$k((p, c), (p', c')) = \frac{|p \cap p'||c \cap c'|}{|p \cap p||c \cap c| + |p' \cap p'||c' \cap c'| - |p \cap p'||c \cap c'|}$$

where $p \cap p'$ indicates the number of fingerprints that kinase domains p and p' have in common. As is shown, the Tanimoto kernel of the pairwise fragments can be efficiently computed using the product of the intersection of kinases and that of compounds. This makes SVM training feasible in an extremely high dimensional space which is spanned by a large number of pairwise fragments.

For the implementation of DCSVMs, we used the LIBSVM library.[48] However, because it was not able to accommodate a large kernel matrix as the input due to lack of memory in the case of >40 000, we used SVM ensembles (see e.g., the works of Caragea et al.[49] and Xu et al.[50]) generated by subsampling. To alleviate the influence of slightly unbalanced numbers of active and inactive pairs, we gave different weights to these two classes via a modified kernel matrix.[51] For SVMs and DCSVMs, the parameter λ in Lauer and Bloch[51] was selected from {0.001, 0.01, 0.1,1, 10, 100}.

Performance Evaluation. To evaluate the performance of our proposed models, internal validation was conducted for data sets A and B, respectively, using random splitting; each data set was partitioned randomly into a training set (90%) and a test set (the remaining 10%). The performance of the internal validation was measured by accuracy = (TP + TN)/(TP + TN + FP + FN), where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. To reliably estimate the performance, we repeated the random splitting 10 times, and the accuracy was averaged to obtain an overall estimate.

Since the numbers of actives and inactives are highly unbalanced for the external data set, we used sensitivity, specificity, and Matthews' correlation coefficient (MCC) as performance measures along with accuracy. Sensitivity and specificity are defined as TP/(TP + FN) and TN/(TN + FP). MCC is given as follows:

$$\text{MCC} = \frac{\text{TPTN} - \text{FPFN}}{\sqrt{(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FP})(\text{TN} + \text{FN})}}$$

Clustering of Compounds. The representation of kinase−inhibitor pairs by a set of pairwise fragments allows us to better interpret kinase cross-reactivity. To this end, we extracted from

Table 1. Statistics of the Data Sets Used for Internal Validation

data set	cutoff	chemical fragments	no. of actives	no. of inactives
A	≤10 μM	RECAP	39 811	23 912
B	<1 μM	RECAP	29 655	36 090
C	≤10 μM	ECFP_6	42 641	25 858
D	<1 μM	ECFP_6	31 513	39 242

kinase profiling data pairwise fragments associated with activity. Here, instead of analyzing all the inhibitors together, we grouped the compounds into clusters and performed analysis of pairwise fragments on a cluster basis. This clustering is based on the premise that structurally similar ligands are presumed to be more closely related with each other in terms of binding profiles. We exploited the affinity propagation (AP) algorithm[52] to perform clustering of the 25 018 compounds on the basis of Tanimoto coefficient (Tc) calculated from the RECAP fragments. To our knowledge, the use of AP in cheminformatics has not been reported earlier, but it has proven to perform better than *k*-means clustering,[52] one of the most common methods for clustering of compounds. There are additional incentives and benefits of exploiting AP in the present study: AP can take pairwise similarities (e.g., Tc values) as the input, and the AP algorithm becomes more efficient when the similarity matrix is sparse, i.e. most of the values are zero, thus allowing the clustering of a large compound data set. Indeed, most of the Tc values between the compounds were zero or close to zero when using the RECAP fragments. By varying the preference value,[52] we determined the number of clusters so that moderately similar compound sets were obtained while maintaining a fitness value as large as possible.

RESULTS

Internal Validation Using SARfari Data Set. Table 1 shows the numbers of kinase−inhibitor pairs with available bioactivity data points used for the internal validation. In the case where the same kinase−inhibitor pair has multiple affinity values, the highest value was employed. The internal validation was performed by random splitting of each data set as described above.

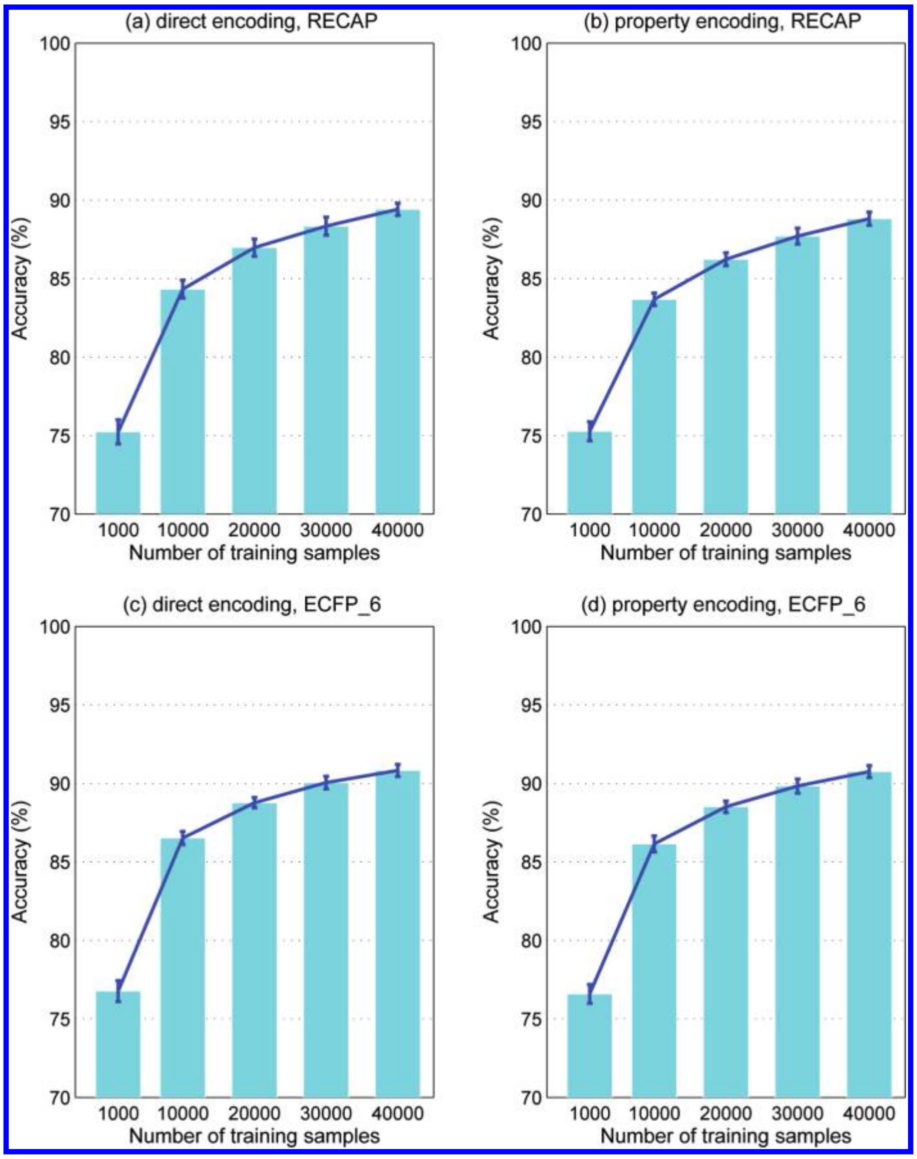


Figure 1. Performance of DCSVMs for different training sample sizes.

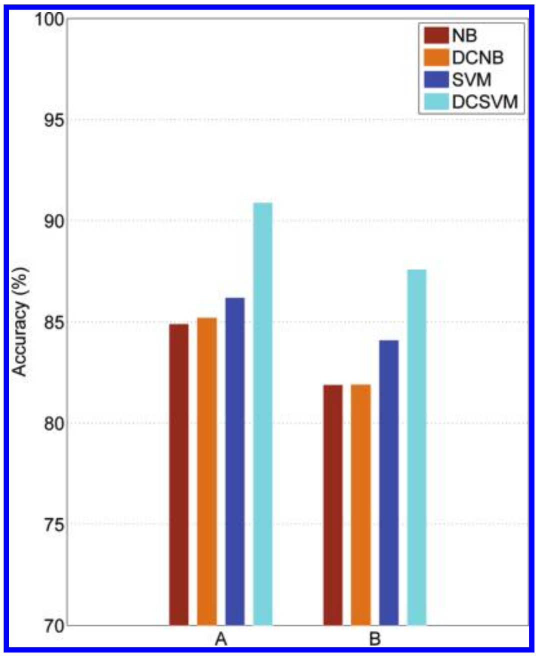


Figure 2. Performance comparison between ligand-based NB, DCNB, ligand-based SVM, and DCSVM for data sets A and B.

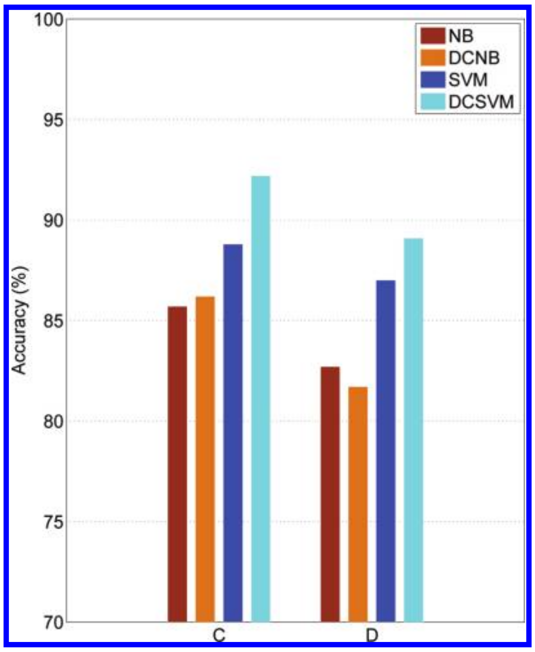


Figure 3. Performance comparison between ligand-based NB, DCNB, ligand-based SVM, and DCSVM for data sets C and D.

Table 2. Performance for Independent Data Sets

data set	cutoff	MCC	accuracy	sensitivity	specificity
SARfari	≤10 μM	0.455	76.9%	80.7%	67.0%
SARfari	<1 μM	0.476	73.9%	75.7%	72.6%
Metz et al.[7]	<1 μM	0.299	81.3%	45.5%	87.2%

Table 3. Performance for Selected Data from Independent Data Sets

data set	cutoff	MCC	accuracy	sensitivity	specificity
SARfari	≤10 μM	0.809	90.8%	91.6%	89.5%
SARfari	<1 μM	0.799	90.1%	88.3%	91.6%
Metz et al.[7]	<1 μM	0.683	90.2%	77.7%	93.0%

We first compared the two encoding methods (direct encoding and property encoding) of amino acid residues using the DCNB model. We obtained the posterior probabilities for all pairwise fragments totaling 38 (36 residues + 2 gaps) × 20 × 15 604 = 11 859 040 for direct encoding and 38 × 21 × 15 604 = 12 451 992 for property encoding and, then, computed the probability of being active and inactive for each tested kinase−inhibitor pair. For data set A, the DCNB models yielded an accuracy of 83.9% (±0.6) by direct encoding and 81.3% (±0.5) by property encoding. The corresponding accuracies for data set B were 81.7% (±0.3) and 78.5% (±0.3), respectively. The total numbers of pairwise fragments with respect to the ECFP_6 fragments were 43 196 120 for direct encoding and 45 355 926 for property encoding. The DCNB models yielded an accuracy of 85.5% (±0.5) by direct encoding and 82.4% (±0.6) by property encoding for data set C. The corresponding accuracies for data set D were 82.2% (±0.2) and 79.2% (±0.3), respectively.

We next compared the performance between the DCNB models and DCSVMs using the direct encoding method. As a result, DCSVMs achieved higher accuracies for all the data sets: 89.5% (±0.3) for data set A, 87.2% (±0.4) for data set B, 90.8% (±0.2) for data set C, and 89.2% (±0.3) for data set D. These results have revealed that DCSVMs significantly outperform the DCNB models.

We further investigated the effect of training sample size on the prediction performance. For this purpose, we used DCSVMs to see how its performance is affected by varying the size of training samples (using data sets A and C). Figure 1 shows that, for both direct encoding and property e, the accuracy of DCSVMs was improved as the training sample size was increased from 1000 to 40 000. Overall, direct encoding showed consistently, albeit marginally, better performance than property encoding, and the ECFP_6 fragments compare favorably to the RECAP fragments.

Comparison with Ligand-Based Approach. It has been reported that the chemogenomics approach is useful when there is no known ligand for a given target, i.e. orphan target, and also shows better performance than the ligand-based approach particularly when the number of known ligands is scarce.[44,45] Thus, it is of interest to compare the performance of the DCNB models (and DCSVMs) based on the direct encoding method with that of the ligand-based NB models (and SVMs) built for individual kinases. To this end, we restricted the analysis to kinases for which ligand-based models were applicable and ≥10 pairs available for both actives and inactives. The numbers of kinases and kinase−inhibitor pairs used were 51 (out of 334) and 42 595 (out of 63 723) for data set A. As shown in Figure 2, the accuracy of the ligand-based NB model was 84.9% (±3.1), which was averaged over all kinases, whereas that of the DCNB model was 85.2% (±0.5). The accuracy was 86.2% (±3.0) for the ligand-based SVM, whereas 90.9% (±0.4) for DCSVM. The large standard deviations of the accuracies of the NB model and SVM were due to the difference in the numbers of known inhibitors across kinases. For data set B, the numbers of kinases and kinase−inhibitor pairs used were 68 (out of 341) and 48 735 (out of 65 745). The accuracies of the NB and DCNB models were 81.9% (±3.7) and 81.9% (±0.4), whereas those of SVM and DCSVM were 84.1% (±3.5) and 87.6% (±0.2) (Figure 2). We also observed similar trends in the prediction performance on data

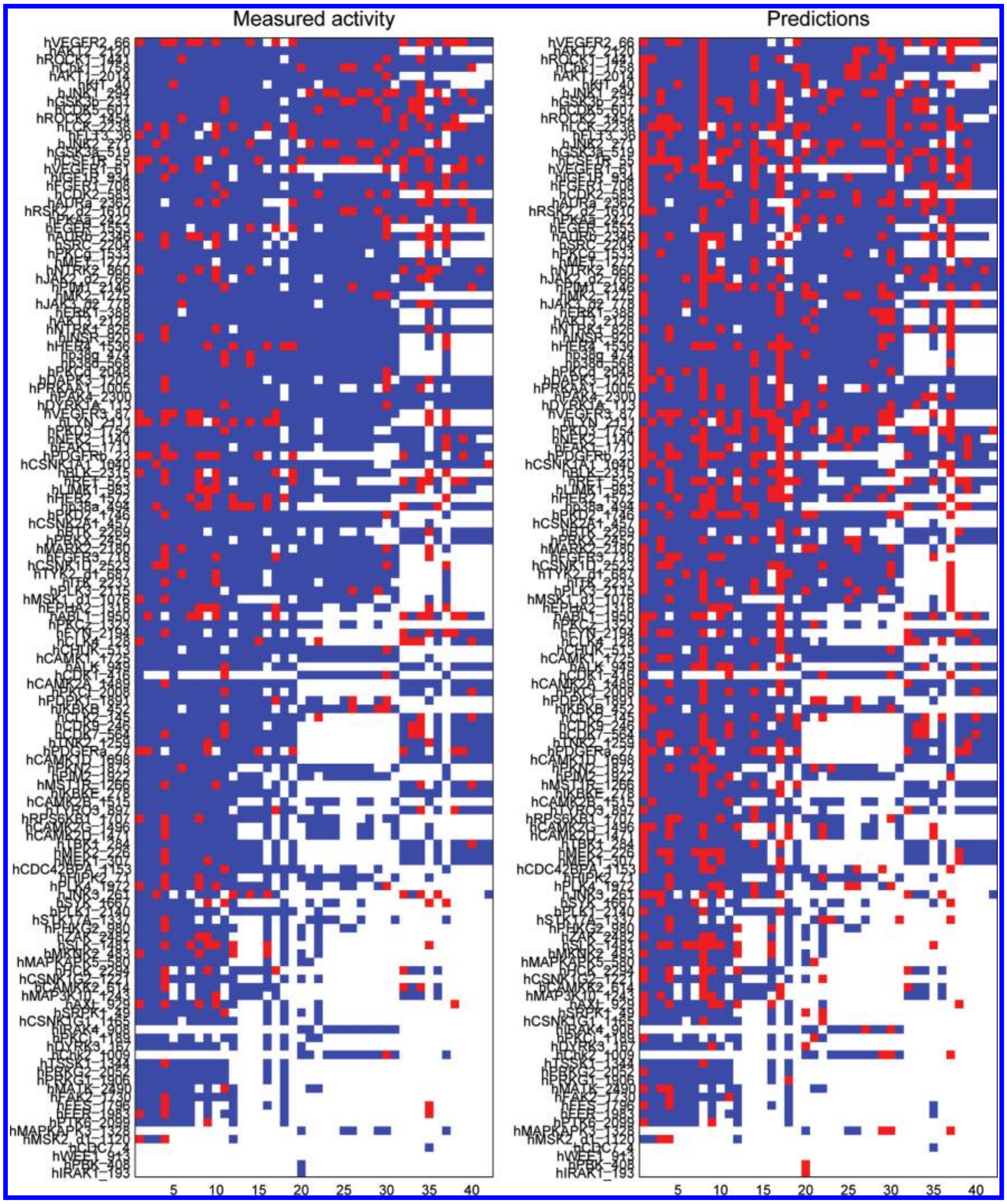


Figure 4. Comparison of measured activity data with their predictions. Actives and inactives are depicted by red and blue, respectively. White cells do not have measured data.

sets C and D: the NB and DCNB models yielded 85.7% (±3.2) and 86.2% (±0.7), whereas SVMs and DCSVMs yielded 88.8% (±2.8) and 92.2% (±0.3) for data set C; the NB and DCNB models yielded 82.7% (±3.7) and 81.7% (±0.5), whereas SVMs and DCSVMs yielded 87.0% (±3.2) and 89.1% (±0.4) for data set D (Figure 3).

External Validation Using Independent Data Sets. We first utilized the latest version (version 5.00) of Kinase SARfari data set for external validation. All of the previous experiments were performed using Kinase SARfari version 3.00, which was downloaded in December 2010, whereas version 5.00 was released in December 2011. We extracted newly added kinase-compound pairs from version 5.00 by excluding the existing pairs in version 3.00 and used them as independent data sets for further validation. The data sets consist of 18 354 pairs (12 388 actives and 5966 inactives) for ≤10 μM and 18 827 pairs (8913 actives and 9914 inactives) for <1 μM, and the number of kinases was 375, increasing from 341.

In the light of the promising performance of DCSVMs with the ECFP_6 fragments, we explored its performance by applying the models to the independent data sets. The model was built using all available pairs of data set C and D, respectively. Since the internal validation indicated that direct encoding was better than property encoding, the former method was used for the prediction models. In addition, we used the optimized parameter that gave the best models in the

Table 4. Comparison of Measured Activity Data with Their Predictions

compound number	compound ID	no. of actives (measured)	no. of actives (predicted)	no. of kinases
1	532	20	84	128
2	3724	6	14	127
3	2521	5	26	127
4	3835	46	46	124
5	3657	5	11	122
6	3438	7	5	122
7	3317	7	13	121
8	766	21	83	118
9	3650	14	23	118
10	3803	25	33	117
11	3211	18	18	116
12	3322	8	17	107
13	3767	1	1	99
14	2137	9	47	99
15	3320	4	4	98
16	3321	8	16	95
17	1204	11	68	91
18	3684	2	9	89
19	3323	11	17	88
20	2918	2	20	86
21	2553	2	13	86
22	2869	4	18	79
23	3028	2	8	78
24	3024	2	5	78
25	2923	4	11	78
26	2917	8	18	78
27	3027	2	6	77
28	3020	3	5	74
29	2915	7	20	69
30	2481	25	37	69
31	3025	2	4	63
32	3855	18	20	62
33	3820	6	8	62
34	3748	26	18	62
35	3738	30	19	62
36	3708	5	8	62
37	714	21	48	60
38	3316	10	13	58
39	3790	12	16	56
40	3136	10	10	56
41	3757	2	1	53
42	3428	1	1	52

internal validation. The resulting models yielded a MCC of 0.455 for $\leq$10 μM and 0.476 for <1 μM. The detailed results are shown in Table 2. As is shown, we could not achieve the same level of performance as in the internal validation. Likely seen in quantitative structure−activity relationship (QSAR) models, it is often difficult to achieve consistently good performance for external validation particularly when models are applied to an independently collected data set, and there can be several reasons for this discrepancy.[53−55] One of the possible reasons in our study could be that the external data set does not fit well to the applicability domain (e.g., the work of Weaver and Gleeson[56]), i.e. ligand-target space of the training set. We therefore hypothesized that the models could achieve better performance for test pairs that lie closer to the applicability domain. To confirm this hypothesis, we selected the pairs out of the independent data sets such that all of their pairwise fragments appear in the training data set. This resulted in 2495 pairs (1478 actives and 1017 inactives) for $\leq$10 μM and 2557 pairs (1101 actives and 1456 inactives). By applying the same models to these pairs, we obtained a MCC of 0.809 for $\leq$10 μM and 0.799 for <1 μM. Table 3 shows the detailed results.

We next used the data set of Metz et al.[7] for further external validation. Of the 172 kinases in the external data set, we were able to map the domain sequences of 139 kinases to the ATP-binding sites of the SARfari data set. The numbers of kinases, inhibitors, and kinase−inhibitor pairs with available bioactivity data points were 138, 1490, and 89 797 (15 613 actives and 74 184 inactives), respectively. Note that MINK was excluded from the analysis because there was no known inhibitor in the SARfari data set. We again applied DCSVMs with the ECFP_6 fragments. The model was built using all available pairs of data set D. The resulting model yielded a MCC of 0.299 for <1 μM. The detailed results are shown in Table 2. We could not achieve the same level of performance as in the internal validation, as previously because the external data set does not fit well to the applicability domain. We therefore selected the pairs such that all of their pairwise fragments appear in the training data set, which resulted in 4898 pairs (972 actives and 3926 inactives). By applying the same model to these pairs, we obtained a MCC of 0.683. Table 3 shows the detailed results.

Activity Predictions to Selectivity Predictions. Although this study has focused on activity predictions for each kinase-compound pair with emphasis on cross-reactivity, rather than selectivity (activity patterns), the expansion of activity predictions to selectivity predictions is feasible. The predicted activities for a range of kinases can provide information about how selective or promiscuous given compounds will be. To illustrate this, we compared measured activity data with their predictions using a portion of the external data set of Metz et al.[7] Specifically, we depicted heat maps for the measured activity and the predictions and checked how selective and promiscuous compounds were predicted by our approach.

Figure 4 and Table 4 contain kinase−compound pairs that were selected out of the 4898 pairs such that the compound activities were measured for more than 50 kinases. As can be seen, our in silico approach tends to predict inactives as actives, particularly for promiscuous compounds (e.g., compounds 532, 766, 2137, 1204, and 714). However, there are several compounds that were predicted as selective (e.g., compounds 3767, 3320, and 3428) and as promiscuous (e.g., compounds 3835, 3211, and 3136). Overall, selectivity predictions remain a formidable challenge for current in silico techniques including our machine learning approach.

Extraction of Activity-Specific Pairwise Fragments. We obtained the posterior probabilities of activity for all possible pairwise fragments based on the RECAP fragments using data set A. Figure 5 shows that activity-specificity varies greatly depending not only on residues and their positions but also on chemical fragments, and this comprehensive map delineates the versatile relationships between residues and fragments. We further performed biclustering of the activity-specificity map on the basis of the probabilities. Interestingly, there are several residues that exhibit activity correlation across a wide array of fragments (Figure 6; region I), whereas some fragments exhibit inactivity correlation across various residues (Figure 6; region II).

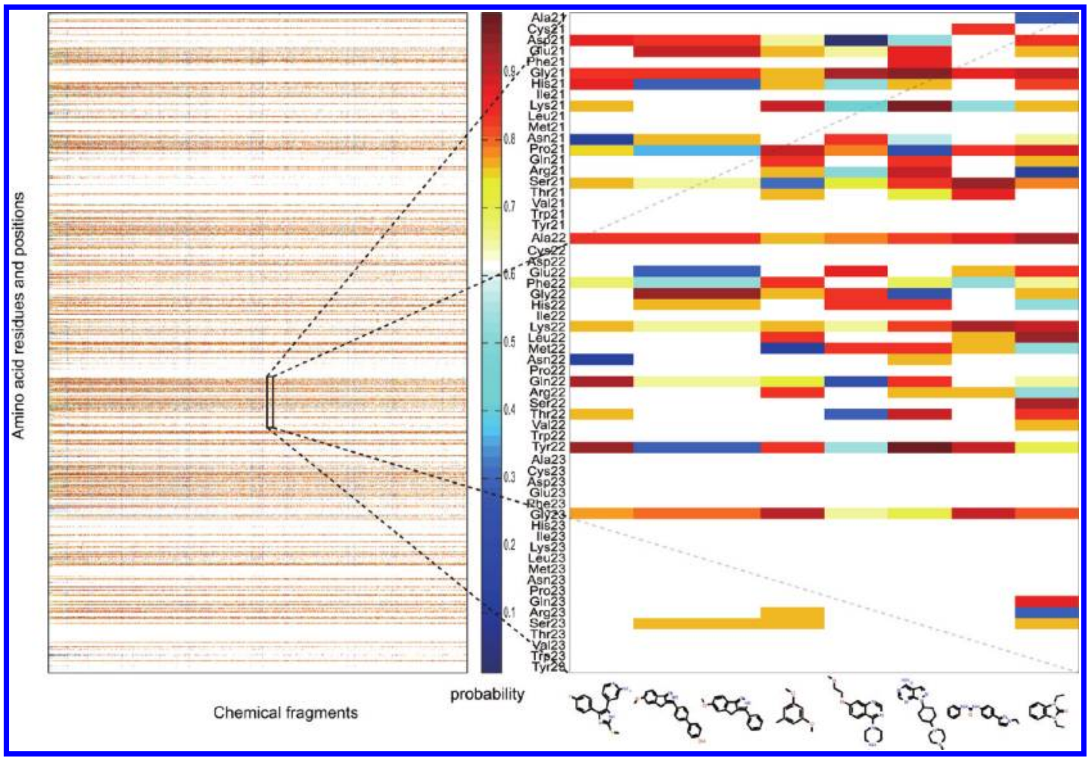


Figure 5. Comprehensive map of activity-specificity. The probabilities for pairs of 20 amino acid residues × 38 positions and 2523 chemical fragments occurring in ≥10 compounds are shown.

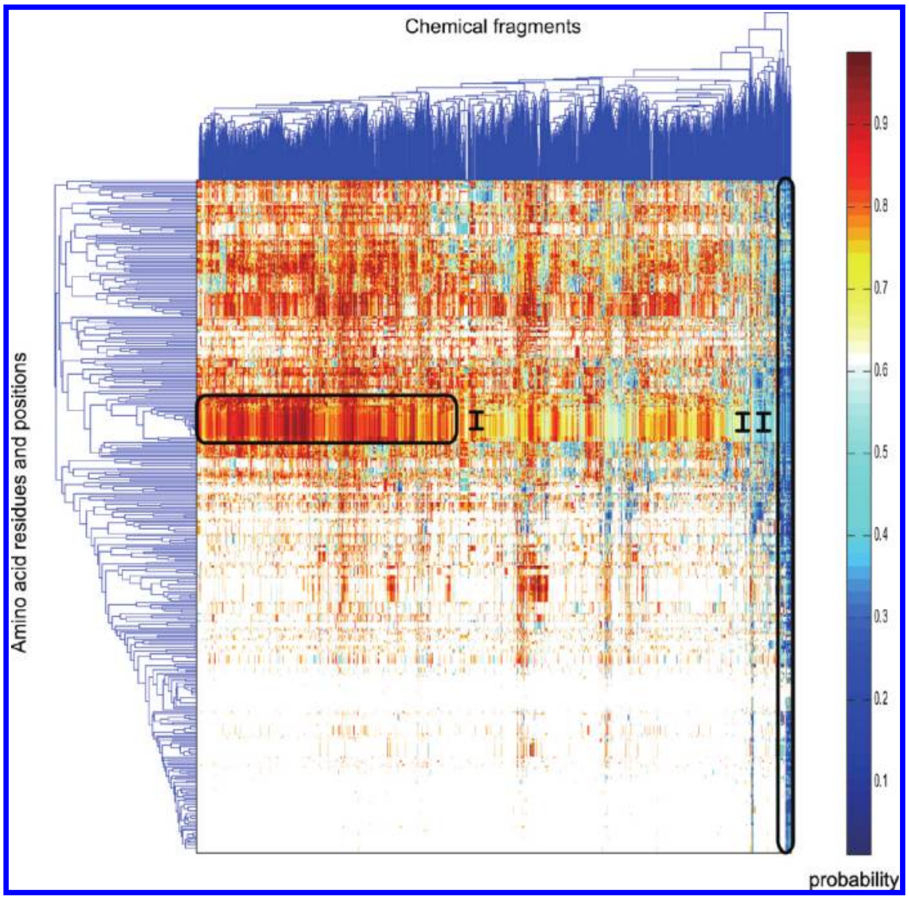


Figure 6. Biclustering of the activity-specificity map.

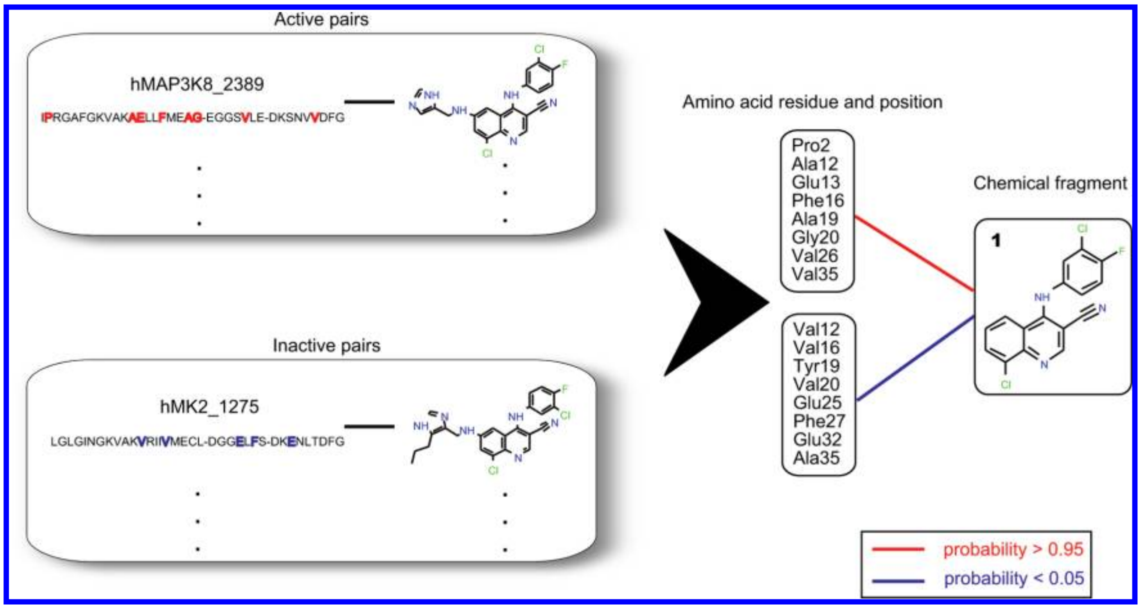


Figure 7. Kinase−compound pairs and extracted pairwise fragments for cluster X (42 active and 46 inactive pairs between 40 compounds and 17 kinases).

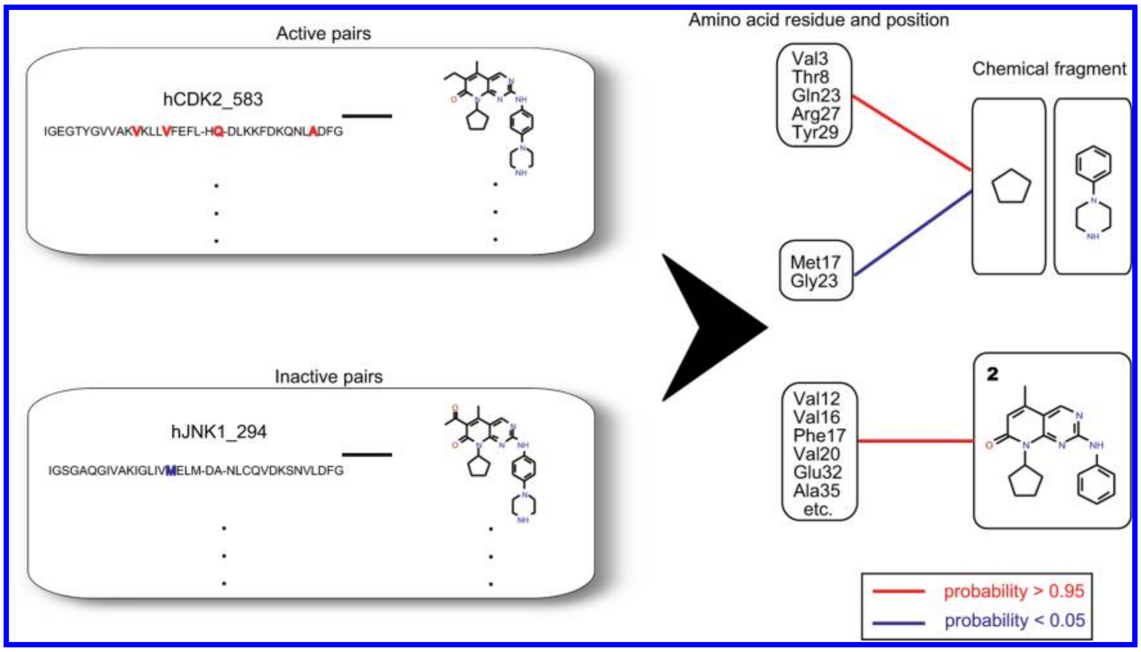


Figure 8. Kinase−compound pairs and extracted pairwise fragments for cluster *Y* (48 active and 37 inactive pairs between 28 compounds and 46 kinases).

Although the comprehensive analysis of the activity-specificity map could yield information about activity-specific pairwise fragments for kinase inhibitors as a whole, focusing on structurally similar compounds may aid in the interpretation of pairwise fragments given that structural similarity is relevant in varying degrees to binding profiles. In view of this, we grouped the SARfari compounds into 3986 clusters exploiting the AP algorithm. Of these, 79 clusters of size ≥20 were selected, and pairwise fragments were extracted by computing posterior probabilities on a cluster basis. We then performed detailed analysis for three clusters (*X*, *Y*, and *Z*) that contained pairwise fragments of both $\Pr(\text{active}|(r_i, f_i)) > 0.95$ and $\Pr(\text{active}|(r_j, f_j)) < 0.05$.

The extracted pairwise fragments and the kinase-compound pairs containing these pairwise fragments are depicted for clusters *X*, *Y*, and *Z* (Figures 7−9). For example, all the compounds in cluster *Z* have fragment 3 in common, and this fragment appears more frequently in active kinase−inhibitor pairs if kinases have Ala34 or Leu35. In contrast, it is less likely so if kinases have Phe17 or Ser22. Thus, compounds containing fragment 3 show preference for kinases having Ala34 or Leu35, and these activity-specific pairwise fragments may partly characterize the cross-reactivity of the compounds.

As shown in Figure 7, kinases having Val12, Val16, Val20, Glu32, or Ala35 are less likely to exhibit activity when paired compounds contain fragment 1. However, these residues are relevant to activity if compounds contain fragment 2 (Figure 8).

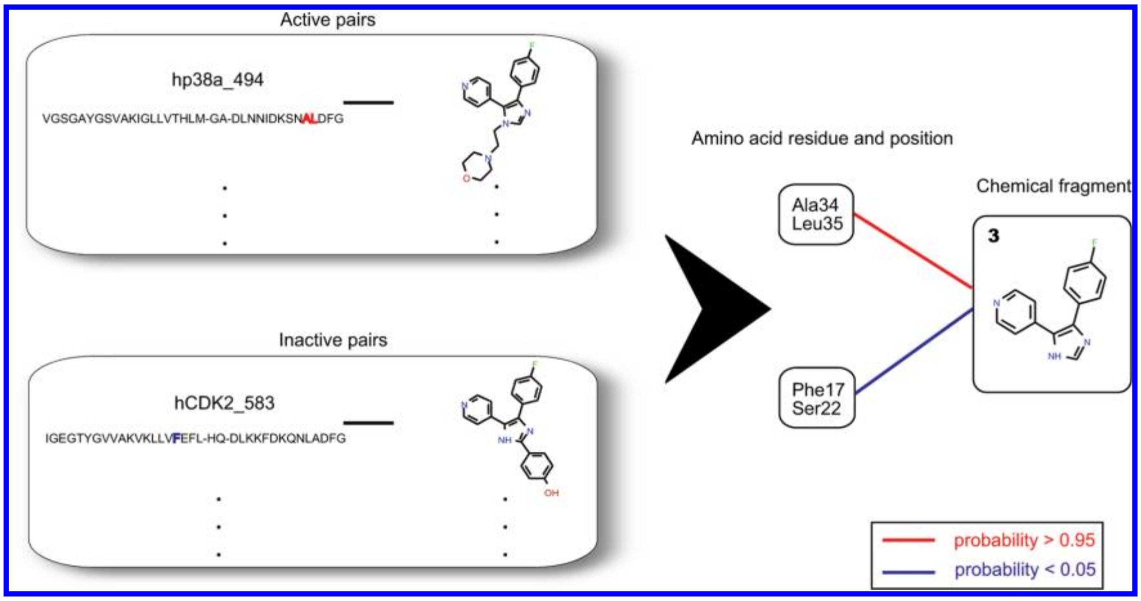


Figure 9. Kinase−compound pairs and extracted pairwise fragments for cluster Z (120 active and 247 inactive pairs between 21 compounds and 283 kinases).

In addition, Phe17 is recurring in inactive pairs if compounds contain fragment 3 (Figure 9), whereas it is recurring rather in active pairs if compounds contain fragment 2 (Figure 8). These results illustrate that whether given compounds exhibit activity against kinases of interest cannot be attributed to either amino acid residue or chemical fragment alone, but residue−fragment pairs are responsible for activity, hence cross-reactivity.

■ DISCUSSION

The experiment for comparing the two encoding methods has shown that direct encoding performs somewhat better than property encoding for both the DCNB models and DCSVMs. This implies that the categorization of residues according to their properties might make discrimination of compound activity less sensitive. The performance of the property encoding method might be improved if more physicochemical properties were encoded, although this could further increase the number of pairwise fragments.

The performance was significantly improved by using DCSVMs instead of the DCNB models. This can be attributed to the fact that DCSVMs can allow for the composite effect of features, while DCNB assumes that the features are mutually independent. This suggests that pairwise fragments are associated with activity in a combinatorial manner. In contrast to the simplicity of DCNB in building the models, however, the computation of DCSVMs is dominated by the training sample size both in memory and time, and the batch learning of SVM was not indeed applicable to a sample size of >~40 000 without contriving to save memory. To cope with this, we used SVM ensembles in the present study. A possible alternative approach to learning from such a large data set may be active learning,[57] which aims to actively select and use only samples that would contribute to prediction. As shown in Figure 1, an increase in the training samples leads to better performance, and hence, learning from large-scale profiling data is emerging as an important, yet formidable task. Therefore, ensemble-based and active learning approaches should play a more crucial role in the near future.[58]

The comparison between our chemogenomics models and the ligand-based models has indicated that DCNB is competitive with NB, and DCSVMs compares favorably to SVMs. The observation that pairs of related kinases and similar compounds are more likely to exhibit similar potency, that is, better correlated, forms the basis of the chemogenomics models. Also note that the training data used were different between the chemogenomics and ligand-based models, and that the better performance of DCSVMs might greatly owe to the observation that the activity of compounds against a kinase of interest can be well correlated with that against the related kinases. Both DCNB and DCSVMs are general models in that these could be applied to any alignable kinase domains, while the application of the ligand-based models is limited to kinases whose ligands are known. Taken together, it was found that DCSVMs with the ECFP_6 fingerprints are particularly effective for activity prediction.

To evaluate the performance of our approach more rigorously, we applied DCSVMs to the independent data sets, which were recently made publicly available. The results have suggested that the model showing high accuracies in internal validation is not able to yield the same level of performance when the tested pairs are apart from the applicability domain, yet the model could be substantially predictive if the pairs fall in the applicability domain or in its vicinity.

The representation of kinase−inhibitor pairs by residues and fragments as dual components could aid in the interpretation of cross-reactivity. We extracted pairwise fragments by computing posterior probabilities for each cluster containing structurally similar compounds. We have illustrated that the preference of given compounds for kinases is better understood by residue−fragment pairs, i.e. pairwise fragments. This suggests that our approach could generate pairwise fragments that serve as selectivity-determining as well as cross-reactivity-determining factors. It should be noted that pairwise fragments that are specific to inactivity are as informative as activity-specific pairwise fragments, because such knowledge can be useful for designing inhibitors that do not target unwanted kinases with the aim of achieving selectivity. Possessing an activity-specific pairwise fragment can be a prerequisite for a given kinase–compound pair to exhibit activity, but does not necessarily suffice by itself. A combination of activity-specific pairwise fragments has significant implications for cross-reactivity on a kinomewide scale, and thus, the question of how to design selective and multitargeted inhibitors reduces to the combination problem of pairwise fragments.

There exist allosteric inhibitors that bind regions outside the ATP-binding site.[4] Although this study has focused on the ATP-binding region of kinases, the same approach could be readily applied to the whole kinase domains once the domain sequences were aligned properly. The number of known allosteric inhibitors is still limited based upon which statistical models are built, but the scalability of our models should enable more comprehensive analysis by widening the range of sequences.

While our approach represents an important step toward the goal of understanding cross-reactivity, some limitations in the proposed models should also be acknowledged. First, because the probabilities are computed from both active and inactive kinase–inhibitor pairs in profiling data, the models require reliable information about inactives as well, which is often less abundant than actives. Second, our approach is based on the assumption that pairwise fragments are mutually independent, hence incapable of identifying features that are activity-specific under the presence of possibly distant features. Last, profiling data collected from different literature sources contain bioactivity data points that are inconsistent in varying degrees due to the difference in binding assay protocols. Such large-scale, noisy, and heterogeneous data prompt the further development of more advanced and robust statistical models.

CONCLUSION

The cross-reactivity issue in kinome has attracted increasing attention. In silico modeling approaches to predict activity for large-scale compound libraries across a wide range of kinases provide a powerful tool for cost-efficient virtual profiling. Among others, the chemogenomics approach based on statistical machine learning that leverages kinase profiling data holds great promise. In this study, we have developed a deconvolution approach to dissecting kinase profiling data, by which kinase–inhibitor pairs are represented by residues and fragments as dual components in order to better interpret and capture kinase cross-reactivity. In particular, we have proposed a dual-component naive Bayes (DCNB) model and dual-component support vector machines (DCSVMs) for activity prediction and performed comprehensive analyses using large-scale data sets. Notably, DCSVMs were found to show promising performance in the internal validation. Although it was difficult to achieve the same level of performance for the independent data sets used, the performance was substantially improved when the applicability domain was properly taken into account. Our proposed approach not only enables activity predictions of given compounds on a kinome-wide scale, but also allows to extract pairwise fragments that are associated with activity. Despite some limitations, our approach represents an important contribution to understanding cross-reactivity between kinases and inhibitors, and the extracted residue–fragments pairs should aid in combinatorial library design as well as lead optimization for selective and multitargeted inhibitors against intended kinases.

AUTHOR INFORMATION

Corresponding Author

*E-mail: niijima@pharm.kyoto-u.ac.jp (S.N.); okuno@pharm.kyoto-u.ac.jp (Y.O.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by grants from Grant-in-Aid for Scientific Research on Innovative Areas “HD Physiology” of the Ministry of Education, Culture, Sports, Science and Technology, Japan; the New Energy and Industrial Technology Development Organization (NEDO) under the Ministry of Economy Trade and Industry of Japan; the Funding Program for Next Generation World-Leading Researchers; and The Okawa Foundation for Information and Telecommunications. Financial support from Ono Pharmaceutical Co., Ltd. is also gratefully acknowledged.

REFERENCES

(1) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1934.

(2) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics–structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.

(3) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839–846.

(4) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28–39.

(5) Fedorov, O.; Müller, S.; Knapp, S. The (un)targeted cancer kinome. *Nat. Chem. Biol.* **2010**, *6*, 166–169.

(6) Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.

(7) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the kinome. *Nat. Chem. Biol.* **2011**, *7*, 200–202.

(8) Sawa, M. Strategies for the design of selective protein kinase inhibitors. *Mini-Rev. Med. Chem.* **2008**, *8*, 1291–1297.

(9) Fabian, M. A.; et al. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

(10) Melnick, J. S.; et al. An efficient rapid system for profiling the cellular activities of molecular libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3153–3158.

(11) Fedorov, O.; Marsden, B.; Pogacic, V.; Rellos, P.; Müller, S.; Bullock, A. N.; Schwaller, J.; Sundström, M.; Knapp, S. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20523–20528.

(12) Karaman, M. W.; et al. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(13) Bamborough, P.; Drewry, D.; Harper, G.; Smith, G. K.; Schneider, K. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. *J. Med. Chem.* **2008**, *51*, 7898–7914.

(14) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689–2700.

(15) Vieth, M.; Erickson, J.; Wang, J.; Webster, Y.; Mader, M.; Higgs, R.; Watson, I. Kinase inhibitor data modeling and de novo inhibitor design with fragment approaches. *J. Med. Chem.* **2009**, *52*, 6456–6466.

(16) Posy, S. L.; Hermsmeier, M. A.; Vaccaro, W.; Ott, K.-H.; Todderud, G.; Lippy, J. S.; Trainor, G. L.; Loughney, D. A.; Johnson, S. R. Trends in kinase selectivity: insights for target class-focused library screening. *J. Med. Chem.* **2011**, *54*, 54–66.

(17) Smyth, L.; Collins, I. Measuring and interpreting the selectivity of protein kinase inhibitors. *J. Chem. Biol.* **2009**, *2*, 131−151.

(18) Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-throughput kinase profiling as a platform for drug discovery. *Nat. Rev. Drug Discovery* **2008**, *7*, 391−397.

(19) Caffrey, D.; Lunney, E.; Moshinsky, D. Prediction of specificity-determining residues for small-molecule kinase inhibitors. *BMC Bioinf.* **2008**, *9*, 491.

(20) Zhang, X.; Fernández, A. In silico drug profiling of the human kinome based on a molecular marker for cross reactivity. *Mol. Pharm.* **2008**, *5*, 728−738.

(21) Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H. Predicting kinase selectivity profiles using Free-Wilson QSAR analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1851−1867.

(22) Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. QSAR models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *J. Chem. Inf. Model.* **2009**, *49*, 1974−1985.

(23) Lapins, M.; Wikberg, J. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinf.* **2010**, *11*, 339.

(24) Ma, X. H.; Wang, R.; Tan, C. Y.; Jiang, Y. Y.; Lu, T.; Rao, H. B.; Li, X. Y.; Go, M. L.; Low, B. C.; Chen, Y. Z. Virtual screening of selective multitarget kinase inhibitors by combinatorial support vector machines. *Mol. Pharm.* **2010**, *7*, 1545−1560.

(25) Bikker, J. A.; Brooijmans, N.; Wissner, A.; Mansour, T. S. Kinase domain mutations in cancer: implications for small molecule drug design strategies. *J. Med. Chem.* **2009**, *52*, 1493−1509.

(26) *Kinase SARfari*; European Bioinformatics Institute, 2011.

(27) Huang, D.; Zhou, T.; Lafleur, K.; Nevado, C.; Caflisch, A. Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis. *Bioinformatics* **2010**, *26*, 198−204.

(28) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W32−W37.

(29) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP−Retrosynthetic Combinatorial Analysis Procedure: powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(30) Schreyer, A.; Blundell, T. CREDO: a protein−ligand interaction database for drug discovery. *Chem. Biol. Drug Des.* **2009**, *73*, 157−167.

(31) RDKit: Cheminformatics and Machine Learning Software, 2010.

(32) *Pipeline Pilot*, version 6.5.1; Accelrys, Inc., San Diego, CA, 2007.

(33) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(34) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(35) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein−small molecule complexes. *Chem. Biol. Drug Des.* **2006**, *67*, 5−12.

(36) Weill, N.; Rognan, D. Development and validation of a novel protein−ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049−1062.

(37) Chen, Y.-F.; Hsu, K.-C.; Lin, S.-R.; Wang, W.-C.; Huang, Y.-C.; Yang, J.-M. SiMMap: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties. *Nucleic Acids Res.* **2010**, *38*, W424−W430.

(38) Klon, A. E. Bayesian modeling in virtual high throughput screening. *Comb. Chem. High Throughput Screening* **2009**, *12*, 469−483.

(39) Bender, A. Chemoinformatics and computational chemical biology. *Methods Mol. Biol.* **2011**, *672*, 175−196.

(40) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(41) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(42) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons, Inc.: New York, 1998.

(43) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626−635.

(44) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(45) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582−592.

(46) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(47) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.

(48) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2001.

(49) Caragea, C.; Sinapov, J.; Silvescu, A.; Dobbs, D.; Honavar, V. Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinf.* **2007**, *8*, 438.

(50) Xu, Y.; Wang, X.-B.; Ding, J.; Wu, L.-Y.; Deng, N.-Y. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *J. Theor. Biol.* **2010**, *264*, 130−135.

(51) Lauer, F.; Bloch, G. Incorporating prior knowledge in support vector machines for classification: a review. *Neurocomputing* **2008**, *71*, 1578−1594.

(52) Frey, B. J.; Dueck, D. Clustering by passing messages between data points. *Science* **2007**, *315*, 972−976.

(53) Maggiora, G. M. On outliers and activity cliffs−Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(54) Johnson, S. R. The trouble with QSAR (or How I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25−26.

(55) Huang, J.; Fan, X. Why QSAR fails: an empirical evaluation using conventional computational approach. *Mol. Pharm.* **2011**, *8*, 600−608.

(56) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−1326.

(57) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(58) Murphy, R. F. An active role for machine learning in drug development. *Nat. Chem. Biol.* **2011**, *7*, 327−330.