JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

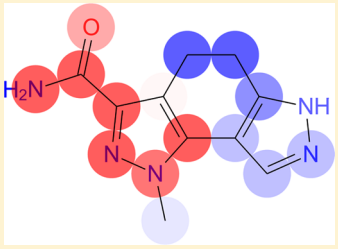
pubs.acs.org/jcim

Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization

Yuling An,[†] Woody Sherman,[†] and Steven L. Dixon[†,*]

[†]Schrödinger, Inc., 120 West 45th Street, New York, New York 10036, United States

ABSTRACT: Numerous regression-based and machine learning techniques are available for the development of linear and nonlinear QSAR models that can accurately predict biological endpoints. Such tools can be quite powerful in the hands of an experienced modeler, but too frequently a disconnect remains between the modeler and project chemist because the resulting QSAR models are effectively black boxes. As a result, learning methods that yield models that can be visualized in the context of chemical structures are in high demand. In this work, we combine direct kernel-based PLS with Canvas 2D fingerprints to arrive at predictive QSAR models that can be projected onto the atoms of a chemical structure, allowing immediate identification of favorable and unfavorable characteristics. The method is validated using binding affinities for ligands from 10 different protein targets covering 7 distinct protein families. Models with significant predictive ability (test set $Q^2 > 0.5$) are obtained for 6 of 10 data sets, and fingerprints are shown to consistently outperform large collections of classical physicochemical and topological descriptors. In addition, we demonstrate how a simple bootstrapping technique may be employed to obtain uncertainties that provide meaningful estimates of prediction accuracy.



INTRODUCTION

The ability to accurately predict biological or chemical properties of compounds, either in a continuous or categorical manner, has been greatly advanced by the introduction of powerful supervised learning techniques, such as partial least squares (PLS) regression,[1] recursive partitioning,[2] naïve Bayes classification,[3] linear discriminant analysis,[4] and support-vector machines (SVM).[5] Perhaps as important as the methods themselves are high dimensional 2D and 3D descriptions of structure[6−11] that have vastly expanded the scope of problems that can be effectively addressed using these techniques.

Once a predictive model has been developed, it is natural to employ that model to direct the synthesis or purchase of subsequent compounds. Although it is possible to apply a given model to a collection of candidate compounds and pursue those with the desired properties of interest, it is generally more desirable to examine which aspects of chemical structure contribute favorably and unfavorably to the model for the predicted property. Such an approach allows for the rational design of subsequent chemical modifications as opposed to enumeration and testing of a much larger number of compounds to arrive at the same endpoint. This goal has been achieved by a variety of approaches, including CoMFA,[10] CoMSIA,[11] 4D QSAR,[12] Phase QSAR,[13] Hologram QSAR,[14] and StarDrop,[15] although only the latter two are independent of 3D structure.

In spite of the numerous advances in supervised learning over the past few decades, benefits to the modeling community at large have been curbed by patents that protect some of the most powerful aforementioned techniques. Although the original patent on CoMFA/CoMSIA has expired, patents are still in effect for SVM[16,17] and for the use of 2D fragments descriptors, or chemical fingerprints, in combination with recursive partitioning[18] and PLS.[19] Curiously, the SVM patents have not been much of a deterrent to its use in practice, as the chemical modeling literature is replete with applications of SVM to QSAR. Fortunately, a highly effective method known as kernel-based PLS (KPLS)[20] remains in the public domain, and its practice in combination with chemical fingerprints is not restricted.

In this paper, we introduce Canvas KPLS, a technique that couples direct kernel-based partial least squares regression[21] with Canvas fingerprints[22−24] to obtain predictive and interpretable 2D QSAR models. The unique nature of Canvas fingerprints, where each fingerprint bit is hashed to a unique chemical fragment, facilitates the direct mapping of each bit in the model back to the atoms in the molecule. This allows a Canvas KPLS model to be readily deconstructed into atomic contributions that can be visualized in terms of favorable and unfavorable structural characteristics on the molecule. Here, we describe the KPLS method and present applications to several congeneric inhibitors for pharmaceutically relevant targets and compare the fingerprint-based approach with a molecular descriptor-based approach. In addition, we present a bootstrapping technique that provides estimations of the uncertainty in the predictions, which are shown to correlate with the accuracy of those predictions over a broad range of biological targets, allowing a researcher to assess when a model will be most predictive.

METHODS

Direct Kernel-Based Partial Least Squares. Direct kernel-based partial least squares (DKPLS)[21] regression uses an orthogonal factorization to construct a low rank

Received: April 26, 2013
Published: July 31, 2013

   dx.doi.org/10.1021/ci400250c | J. Chem. Inf. Model. 2013, 53, 2312−2321

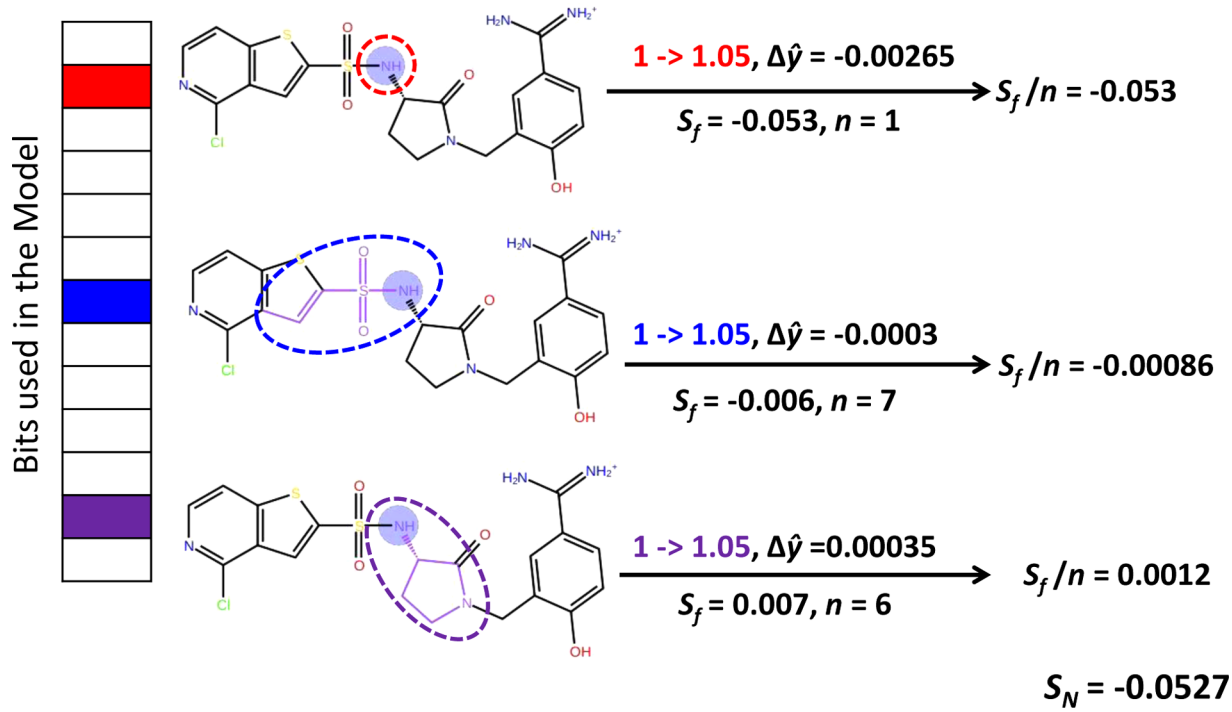


Figure 1. Illustration of the effect on the predicted activity exerted by a sulfonamide nitrogen atom (the attached hydrogen atom is ignored). For each fragment f included in the model by way of its fingerprint bit, $\Delta\hat{y}$ is the change in predicted activity that occurs when the "on" bit value of 1 is perturbed to 1.05. The sensitivity S_f to the presence of fragment f is given by $\Delta\hat{y}/0.05$, and this value is distributed equally over the atoms in the fragment. The overall effect S_N exerted by the sulfonamide nitrogen atom is the sum of its contributions to the various fragments containing that nitrogen.

approximation of the kernel matrix **K**, and then employs this approximation to compute the final regression function. DKPLS is highly efficient relative to eigenvector methods, and the low rank approximation it produces is well suited for regression as it tends to yield more robust models. Theoretical details of DKPLS and a comparison to standard KPLS can be found in ref 21. Here, we focus on how the method is adapted for use with Canvas hashed fingerprints, visualization of QSAR models, and calculation of uncertainties in the model predictions. Henceforth, we refer to the method employed in this work as KPLS even though we are using the direct formalism.

Before constructing a QSAR model from a training set of compounds, an initial pool of independent x variables is assembled from the union of fingerprint bits that are "on" in at least one compound in the training set. Each x variable is thus a binary valued descriptor (0/1) representing a particular chemical fragment that is found at least once. A given x variable is discarded if its mean value is seven or more standard deviations away from the mean of all x variables in the pool. This process succeeds in removing fingerprint bits that are nearly always off (a mean very close to zero) or nearly always on (a mean very close to one). Additional x variables are eliminated by discarding one variable from each pair that correlates more strongly than 0.95. The remaining x variables are then autoscaled (i.e., mean of zero and standard deviation of one), with all scaling factors being saved for application to any subsequent test set data.

For a pair of autoscaled variables, denoted by the column vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, a Gaussian kernel matrix is defined as follows

$$\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2) \quad (1)$$

The nonlinearity parameter σ can be *tuned* to optimize the accuracy of leave-n-out (LNO) predictions, but we found that doing so resulted in σ values that fluctuated significantly with the choice of random seed for selecting LNO subsets. However, model predictions were observed to be essentially insensitive to σ when its value was 20 or larger, so σ was held fixed at 20 for all computations. Training and test set kernel matrices were centered as described in ref 21, and KPLS models containing a maximum of five latent factors were constructed.

Resolving QSAR Models into Atomic Effects. For a given structure i, the predicted activity $\hat{y}_i$ depends only on the fragments f_j in that structure, which correspond to independent variables x_j utilized in the model. Each such variable will be assigned a value of 1 for that structure because the presence of fragment f_j turns "on" the associated bit. To determine whether fragment f_j tends to increase or decrease predicted activity in structure i, a small perturbation of 0.05 is added to x_j (i.e., $1 \rightarrow 1.05$), and the activity is recomputed, affording a change $\Delta\hat{y}_{ij}$ in the predicted activity. The *sensitivity* S_{fj} to fragment f_j is then computed as $\Delta\hat{y}_{ij}/0.05$. It is worth noting that while the perturbation of 0.05 is an arbitrary value, no significant effect on sensitivity was observed for perturbations in the range 0.01−0.05.

Assuming that each atom in fragment f_j contributes equally to S_{fj}, dividing this quantity by the number of heavy atoms n_j in the fragment affords *per atom* sensitivities. An overall sensitivity S_a to the presence of a particular atom a is obtained by accumulating contributions from all fragments f_j that contain atom a

$$S_a = \sum_{j:a\in f_j} S_{f_j}/n_j \quad (2)$$

dx.doi.org/10.1021/ci400250c | J. Chem. Inf. Model. 2013, 53, 2312−2321

The atomic sensitivity S_a is thus a first order approximation to the effect that atom a exerts on predicted activity, and it provides a convenient means of visualizing favorable and unfavorable characteristics of chemical structure, as shown in Figure 1. Throughout this work, atoms that increase (i.e., improve) predicted activity are shaded red, whereas atoms that decrease it are shaded blue.

Estimating Prediction Uncertainties. Bootstrapping is a well-established procedure for estimating the variability of statistical parameters that are derived from a sample of a population. For a collection of M individuals, bootstrapping involves drawing a random sample of M members of that collection, with replacement, such that some individuals will be represented more than once. This allows a distribution of parameter values to be derived from samples that are equal in size to the original collection, which preserves various statistical conditions (e.g., degrees of freedom) that would otherwise differ if the sample size were smaller.

In the case of Canvas KPLS, bootstrapping is applied to the training set, and a model is built from each bootstrapped sample. A total of N models are built, and each model is applied to a given test set compound to obtain a distribution of predictions. The uncertainty $\delta\hat{y}_i$ in the prediction for compound i is then computed as the standard deviation of that distribution, treating the prediction obtained from using the full training set, $\hat{y}_i$, as the mean of the distribution

$$\delta\hat{y}_i = \left[\frac{\sum_{k=1}^{N}(\hat{y}_i^k - \hat{y}_i)^2}{N-1}\right]^{1/2} \quad (3)$$

For computational expediency, N was set to 10 for all work presented here. The bootstrapping procedure is illustrated in Figure 2.

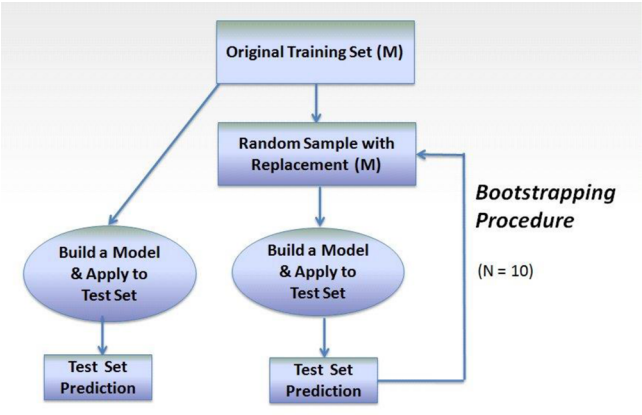


Figure 2. Bootstrapping procedure used to estimate the uncertainty in predicted activities. M is the training set size, and N is the number of bootstrapped samples.

This approach involves no direct attempt to evaluate the model applicability domain,[25−28] so, for example, the similarity of a test set structure to members of the training set is not considered when estimating the uncertainty. Instead, we are considering the variability of predictions that would be expected to occur from using different training sets drawn from the same population. Parameters of the model that are particularly sensitive to the choice of training set will therefore govern the variability of the predictions. In turn, a test set compound that contains features that are keyed to highly variable parameters will tend to be predicted with a high degree of uncertainty. Thus, even a structure that is highly similar to one or more members of the training set may receive a large estimated prediction uncertainty. Model applicability domain methods will generally not flag these compounds as potential outliers because they appear to fall within the trusted region.

Estimation of uncertainties does come with a price because it requires rebuilding and reapplying the model 10 times. Thus if t_{train} is the time required to build a model and t_{test} is the time required to apply it to a test set, the total time will be $10 \times t_{train} + 11 \times t_{test}$. On a modern Linux workstation, both t_{train} and t_{test} are slightly less than 10 s for training and test sets of 100 compounds apiece. The times increase linearly with the sizes of the training and test sets. Memory requirements are quite modest, so scaling up to thousands of compounds presents no practical challenges other than relatively long run times.

RESULTS AND DISCUSSION

Data Sets. A total of 1200 ligands with published binding affinity data[29−38] were used in this study. The ligands cover 7 protein families and 10 different proteins with a range of 73−203 ligands per protein. Each collection was randomly divided into training and test sets, enforcing a 3:1 training:test ratio. All affinities were converted to molar concentrations, and activities were computed as the negative base 10 logarithm of those concentrations. Table 1 displays the size, average activity, and standard deviation in activity for each training and test set.

Before building QSAR models, we first studied the relationship between structural similarity and activity differences for pairs of training and test set compounds. Using Canvas dendritic fingerprints and Tanimoto similarities, each test set compound was compared to its nearest neighbor in the training set. The structural similarity of that pair was then plotted against the difference in activity. As shown in Figure 3, there is no clear relationship between similarity and activity differences. The majority of pairs are highly similar (between 0.8 and 1), yet many of these compounds differ in activity by an order of magnitude or more. This should come as no surprise because a small modification to a structure, such as changing the position of a single substituent from *ortho* to *meta*, can drastically affect the ability to bind but has little affect on the 2D similarity. These observations run somewhat counter to the notion that similar compounds tend to exhibit similar activities and suggest that only a relatively small number of bits in a chemical fingerprint are likely to be relevant for predicting activity.

Choice of Fingerprints. As noted previously, one of the major advantages of KPLS is its ability to utilize chemical fingerprints for model building without infringing on patents. Canvas provides eight types of fingerprints, which come with a host of atom typing schemes. Rather than generating results for all data sets using all possible types of fingerprints, we randomly selected three data sets from Table 1 and built KPLS models for each set using four fingerprint types (dendritic, linear, molprint2D, and radial) that were previously shown to perform well in virtual screening studies.[23] A maximum of five latent factors were utilized in each model, but this number was automatically reduced, as necessary, to prevent the standard deviation of regression (SD) from dropping below 0.4. Unless otherwise noted, results correspond to this method of choosing the number of latent factors.

Table 1. Summary of Data Sets Used in This Paper

				size		training set activity		test set activity	
PDB Code	receptor family	abbrev.	ref.	training	test	average	SD	average	SD
1fvt	cyclin dependent kinase 2	cdk2	29	72	23	7.632	1.019	7.831	0.842
2bkz	cyclin dependent kinase 2	cdk2	30	55	18	6.851	1.055	6.954	0.995
2e9u	CHK1 checkpoint homologue	chk1	31	65	21	7.761	0.764	7.632	0.847
1f0r	factor Xa	fxa	32	75	24	7.031	0.944	6.993	0.897
1z6e	factor Xa	fxa	33	85	28	8.857	1.529	8.748	1.596
2ccu	heat shock protein 90	hsp90	34	60	20	6.661	1.037	6.704	1.020
3kfc	liver X receptor beta	lxrb	35	110	36	7.789	0.850	7.846	0.807
1r5g	methionine aminopeptidase 2	metap2	36	153	50	6.887	0.925	6.904	0.905
2adu	methionine aminopeptidase 2	metap2	37	83	27	8.213	1.020	8.277	0.985
1mue	thrombin	throm	38	147	48	8.609	1.309	8.627	1.282

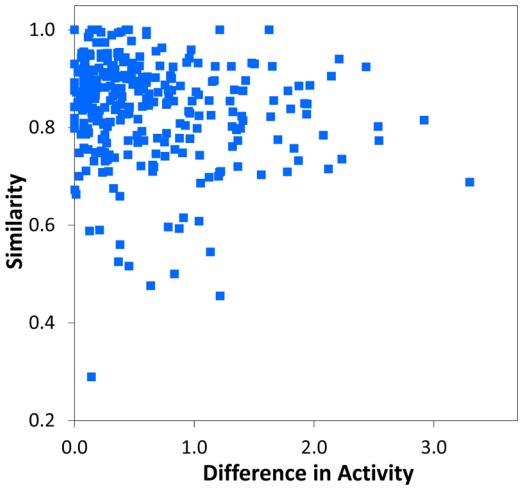


Figure 3. Tanimoto similarities and activity differences between each test set compound and its nearest neighbor in the training set.

Tables 2A and 2B contain training set and test set statistics, respectively, for models built using the four top performing fingerprints from ref 23. Throughout this paper, R^2 is the coefficient of determination for the training set fit, while Q^2 is an analogously computed statistic for the test set predictions

$$R^2 = 1 - \frac{\sum_i^{\text{training}} (y_i - \hat{y}_i)^2}{\sum_i^{\text{training}} (y_i - \overline{y}_{\text{training}})^2} \quad (4)$$

Table 2A. Training Set R^2 Values for a Variety of Canvas Fingerprint Types

data set	dendritic	linear	radial	Molprint2D
cdk2_1fvt	0.877	0.868	0.860	0.825
fxa_1f0r	0.870	0.836	0.862	0.841
hsp90_2ccu	0.902	0.929	0.876	0.872
average	0.883	0.878	0.866	0.846

Table 2B. Test Set Q^2 Values for a Variety of Canvas Fingerprint Types

data set	dendritic	linear	radial	Molprint2D
cdk2_1fvt	0.585	0.500	0.404	0.398
fxa_1f0r	0.514	0.465	0.457	0.553
hsp90_2ccu	0.872	0.902	0.751	0.774
average	0.657	0.622	0.537	0.575

$$Q^2 = 1 - \frac{\sum_i^{\text{test}} (y_i - \hat{y}_i)^2}{\sum_i^{\text{test}} (y_i - \overline{y}_{\text{test}})^2} \quad (5)$$

Here, $\overline{y}_{\text{training}}$ and $\overline{y}_{\text{test}}$ are the average observed activities within the training and test sets, respectively. Note that the least squares fitting procedure guarantees that the model will be at least as accurate as predicting the mean $\overline{y}_{\text{training}}$ for every training set compound, so R^2 will always be non-negative. The same is not always true of Q^2, however, because the numerator in eq 5 may exceed the denominator if the model performs sufficiently poorly on the test set, and/or if the variance in the test set activities is sufficiently small.

Tables 2A and 2B show that dendritic fingerprints outperform the others, on average, for both training and test sets. Interestingly, the advantage of dendritic fingerprints in diversity analysis and hole filling applications was also recently demonstrated,[39] supporting the view that this fingerprint type is versatile and robust for structural characterization. For this reason, dendritic fingerprints were used in all subsequent work.

It is worth noting that employing the number of occurrences of each fragment in a structure did not offer any advantage over a simple 1/0 approach. For example, models built from counts of dendritic fingerprint fragments yielded an average Q^2 of 0.598 for the data sets in Tables 2A and 2B , which is significantly lower than the value of 0.657 obtained from the 1/0 scheme. Although somewhat counterintuitive, this may simply be a consequence of ascribing an equal, additive effect to each occurrence of a given fragment, regardless of where that fragment is found in the structure. While the 1/0 approach certainly does not distinguish where the fragment or fragments are located, it does impose a strict bound on the impact of multiple erroneous fragments.

One final point is that all the statistics reported here are dependent on the particular choice of training and test set compounds. Generally speaking, though, the KPLS method in combination with Canvas fingerprints is not overly sensitive to the division of a given data set. For example, building 10 additional dendritic-based models from 10 different random 3:1 splits of the cdk2_1fvt data set yielded Q^2 values ranging from 0.514 to 0.648, with a mean of 0.600, which is not far from the Q^2 value of 0.585 reported in Tables 2A and 2B.

Fingerprints vs Ordinary Descriptors. To provide a standard of comparison, KPLS models were also built from the collection of over 600 physicochemical and topological descriptors (henceforth referred to as "descriptors") that may be calculated using Canvas. The training/test set compositions and all model building options were the same for both types of

dx.doi.org/10.1021/ci400250c | J. Chem. Inf. Model. 2013, 53, 2312−2321

Table 3. Comparison of test set predictions for KPLS models built from dendritic fingerprints, physicochemical/topological descriptors, and a combination of the two. Results are reported for choosing the number of factors according to the rule SD ≥ 0.4 and for the maximum of 5 factors

	fingerprints		descriptors		fingerprints + descriptors	
data set	Q^2(#factors)[a]	Q^2(5)	Q^2(#factors)[a]	Q^2(5)	Q^2(#factors)[a]	Q^2(5)
cdk2_1fvt	0.585 (3)	0.635	0.058 (5)	0.058	0.365 (3)	0.409
cdk2_2bkz	0.772 (2)	0.787	0.490 (4)	0.578	0.745 (3)	0.792
chk1_2e9u	0.394 (3)	0.403	0.361 (5)	0.361	0.455 (3)	0.502
fxa_1f0r	0.514 (3)	0.409	0.451 (5)	0.451	0.556 (3)	0.506
fxa_1z6e	0.594 (5)	0.594	0.662 (5)	0.662	0.679 (5)	0.679
hsp90_2ccu	0.872 (2)	0.880	0.674 (4)	0.638	0.822 (2)	0.853
lxrb_3kfc	0.383 (5)	0.383	0.132 (5)	0.132	0.360 (4)	0.286
metap2_1r5g	0.701 (4)	0.747	0.424 (5)	0.427	0.677 (4)	0.735
metap2_2adu	0.227 (4)	0.238	−0.327 (5)	−0.327	0.080 (4)	0.069
throm_1mue	0.589 (5)	0.589	0.570 (5)	0.570	0.737 (4)	0.723
average	0.563 (3.6)	0.567	0.349 (4.8)	0.355	0.548 (3.5)	0.555

[a]The model utilized the largest number of factors, up to a maximum of 5, that resulted in a standard deviation of regression which was no smaller than 0.4.

independent variables. Test set statistics for models built from fingerprints and descriptors are compared in Table 3, and scatter plots for fingerprint models are shown in Figure 4. For 9 out of 10 data sets, fingerprints yielded more accurate test set predictions than descriptors. Only in the case of fxa_1z6 did the descriptors perform better, although both sets of results were respectable, with Q^2 values in the neighborhood of 0.6. On average, models built from fingerprints significantly outperform their descriptor-based counterparts (0.57 versus 0.36 for the five-factor models), and most importantly, the average Q^2 value exceeded 0.5, which means the fingerprint models accounted for more than half of the variance in the test set activities.

When the number of latent factors was chosen based on the rule SD ≥ 0.4, fingerprint models tended to contain a smaller number of factors than descriptors models. This is because fingerprint bits are generally more orthogonal to each other than whole molecule descriptors and because the bits contain highly specific substructure information relevant to ligand binding, which is usually not as effectively encoded by ordinary descriptors. Hence, a given level of fitting is achievable with a smaller number of latent factors. The superiority of fragment-based descriptors (i.e., fingerprints) for explaining structure−activity data is well-known,[7,8,40] and our findings are further confirmation of this.

Using a combination of fingerprints and descriptors yielded predictions that, on average, were intermediate in accuracy between the two approaches, although much closer to fingerprints alone. While one might anticipate this result, the combination models actually performed better than pure fingerprints in several cases, even when descriptors alone performed poorly. Perhaps the most surprising results were for throm_1mue, where the combination model yielded a Q^2 of 0.74, which was significantly higher than either the fingerprint model (Q^2 = 0.59) or the descriptors model (Q^2 = 0.57). This suggests that in the case of throm_1mue, the two sets of x variables are providing information that is sufficiently orthogonal to yield a markedly synergistic effect.

It is natural to wonder whether a QSAR method that performs well on congeneric series, as examined here, will perform equally well on diverse data sets. Indeed, a chemical series that spans distinct scaffolds is expected to be more challenging because the QSAR method must reconcile gross differences in chemical structure as well as the possibility of different binding modes. One rigorous way to test this hypothesis is to simply combine all of the training sets in this study and build a single model of ligand−receptor binding that covers all targets. Whether such a model has any physical significance is obviously debatable, but if it is predictive across the combined test sets, the ability of the QSAR method to tolerate a diverse series is certainly implied. The results of this exercise are shown in the "Combined" plot in Figure 4. Here, a single five-factor dendritic model was built from all 905 training compounds across the 10 data sets and then applied to the combined test set of 295 compounds. This yielded a Q^2 value of 0.686, which is considerably higher than the average statistic of 0.567 obtained by averaging the five-factor Q^2 values over the 10 data sets (Table 3). It would therefore appear that fingerprint-based KPLS can be effective on diverse data sets.

Model Visualization. Although it is often possible to infer a basic structure−activity relationship (SAR) through careful analysis of affinities from a well-designed medicinal chemistry series, certain intricacies and subtleties of the relationship are difficult to elucidate without the aid of a reliable statistical model and a visual projection of that model onto chemical structures. Furthermore, being able to visualize favorable and unfavorable characteristics of structure helps bridge the gap between simply knowing which features are important for activity and which new compounds to synthesize or acquire as part of a lead optimization effort.

A visual representation of atomic effects for Canvas KPLS models built from dendritic fingerprints are shown in Figure 5 for compounds with weak, moderate, and strong affinities against three of the targets analyzed in this study. Atoms that increase predicted activity are colored red, whereas atoms that decrease it are blue. Color intensity reflects the strength of the effect.

Of particular note is that the strength and algebraic sign of the atomic effect at a particular position in the chemical backbone can vary in progressing from weak, to moderate, to strong binders. This is a consequence of certain fingerprint fragments being comprised of atoms from the fixed backbone and atoms from a variable region. Exchanging one substituent for another can therefore result in the addition and removal of bits that span the fixed and variable regions. If the net effect is a gain in "good" bits, a backbone atom associated with those bits

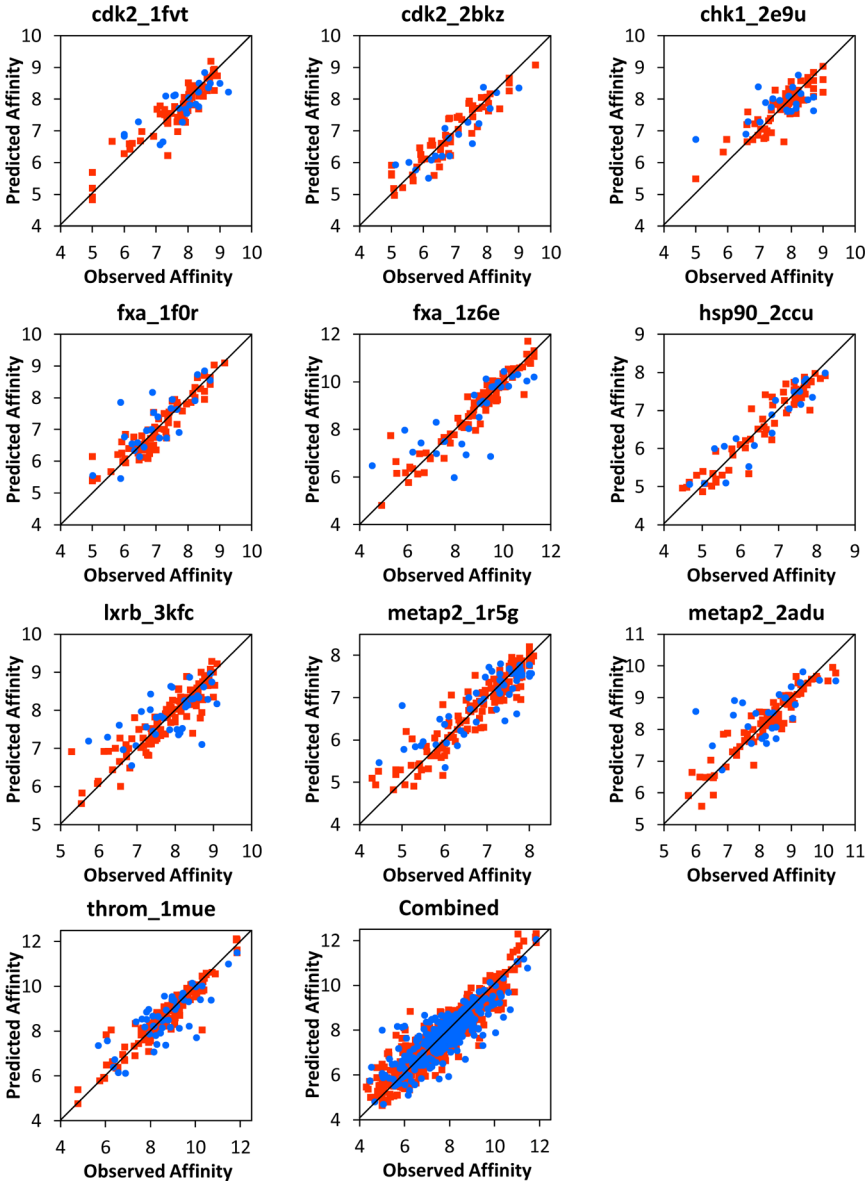


Figure 4. Predicted and observed affinities for training set (red) and test set (blue) compounds using KPLS models constructed from dendritic fingerprints. A maximum of five latent factors was used in each model subject to the restriction SD ≥ 0.4. The "Combined" plot contains results for a single five-factor KPLS model built from a 905 compound training set created by combining all training compounds from all 10 data sets.

may turn from blue to red. Whereas it is customary to view the backbone as exerting a constant effect, the model suggests that parts of the backbone may be irrelevant or even unfavorable unless the correct substituent is attached.

The process of exploiting these visual representations to guide the synthesis or purchase of new compounds varies from one compound series to the next, but the general protocol is to identify the favorable characteristics of moderate and strong binders and infer transformations that would result in unique combinations of favorable features. There is of course nothing new in such an approach, but it is made much easier by a visual embodiment of the SAR.

Figure 6 contains a schematic illustration of the optimization process for inhibitors from the cdk2_1fv2 set. Compounds A and B are highly similar and are predicted to have equally moderate affinity. A key difference is the replacement of an imine linkage by a methylated vinyl, which evidently results in a favorable interaction (see circles in Figure 6). This finding is not obvious from the two structures and their affinities alone. Structure C, a fairly strong binder, contains a nonmethylated vinyl linkage, a favorable oxazole substituent, and an apparently unfavorable sulfolene in place of the phenyl sulfonamide. From these pieces of information, one might reasonably infer that a compound with the methylated vinyl *and* the oxazole would be a stronger binder than any of the other three, and compound D confirms that this is in fact the case.

Reliability of Prediction Uncertainties. A prediction may be inaccurate for any number of reasons, and it is not realistic to expect that a technique exists that can accurately estimate on a consistent basis the size of prediction errors. Even providing a useful upper or lower bound on the error is extremely challenging, and any meaningful analysis of prediction uncertainties must involve examination of overall trends rather than individual examples.

Here, we focus on distributions of absolute prediction errors for compounds that are assigned small prediction uncertainties,

dx.doi.org/10.1021/ci400250c | J. Chem. Inf. Model. 2013, 53, 2312−2321

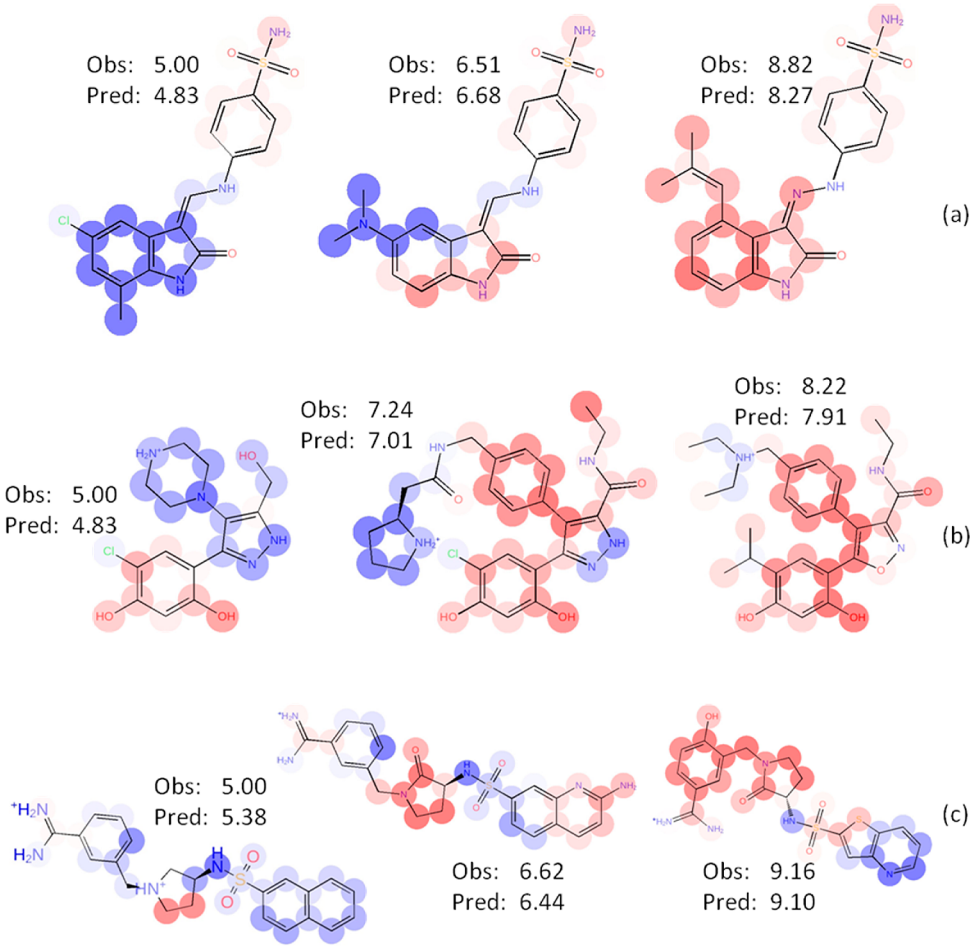


Figure 5. Visualization of atomic effects for Canvas KPLS models built from dendritic fingerprints. Weak (first column), moderate (second column), and strong (third column) binders are included for (a) cdk2_1fv2, (b) hsp90_2ccu, and (c) fxa_1f0r.

moderate prediction uncertainties, etc. Accordingly, all 295 test set compounds from the 10 data sets analyzed in this study were divided into five subsets according to the value of their prediction uncertainty: [0, 0.21], [0.21, 0.3], [0.3, 0.39], [0.39, 0.51], and [0.51, 1.25]. Interval widths are nonuniform because the subsets were chosen so as to obtain equal populations in each bin. A cumulative probability distribution of absolute prediction errors was then constructed for each subset, Figure 7.

It is evident from Figure 7 that compounds with the smallest uncertainties tend to have smaller prediction errors. In fact, 50% of the compounds with uncertainties below 0.21 exhibit an absolute error of less than 0.25 log units, and 90% of them are predicted to an accuracy of 0.75. By contrast, only 15% of compounds with the greatest uncertainties (above 0.51) are predicted to within 0.25 log units, and only 65% are predicted with an accuracy of 0.75 log units or better. For compounds in this most uncertain category, one must go to an error of 1.75 in order to encompass 90% of this group. Distributions for compounds assigned to the intermediate intervals are not clearly distinguishable from a distribution that includes all 295 compounds.

These results suggest that by focusing on predictions with low uncertainties, one can significantly increase the probability of a highly accurate prediction and greatly reduce the possibility of a gross misclassification. Avoiding compounds with uncertainties above 0.51 is not as effective a technique for singling out poor predictions because some of these compounds are in fact predicted accurately. Nevertheless, the most egregious errors are associated with compounds in this group.

Although the uncertainty ranges in this study were derived from 10 different data sets spanning 7 distinct target families, there is no reason to believe that the ranges are universal. However, if a test set is held out for validation purposes, it is straightforward to sort the test set predictions by increasing uncertainty and set a threshold based on, e.g., the first 20%.

■ CONCLUSIONS

The combination of direct kernel-based PLS with Canvas fingerprints has been shown to produce highly predictive QSAR models for a wide variety of target families. On average, these models account for more than half the variance in test set activities, substantially exceeding the predictive ability of models created from large numbers of ordinary physicochemical/topological descriptors. The unique nature of Canvas hashed fingerprints allows KPLS models to be decomposed into atomic effects, which provides a convenient way of visualizing favorable and unfavorable aspects of chemical structures, thereby elucidating the SAR with a high degree of detail. In addition, model visualization can provide clues for modifying a structure to optimize potency, an approach that may be preferred to simply applying the model to collections of compounds. Finally, we have shown how a simple boot-

dx.doi.org/10.1021/ci400250c | J. Chem. Inf. Model. 2013, 53, 2312−2321

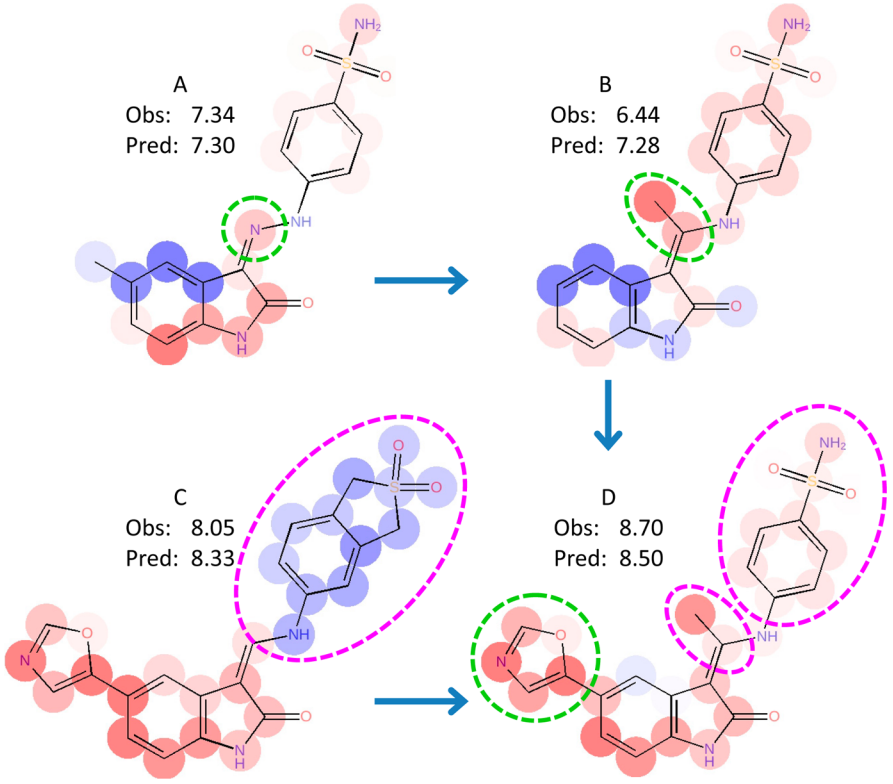


Figure 6. Illustration of how KPLS model visualization may be used to infer structural transformations that are likely to improve affinity. Color-coded circles indicate key differences between pairs of structures.

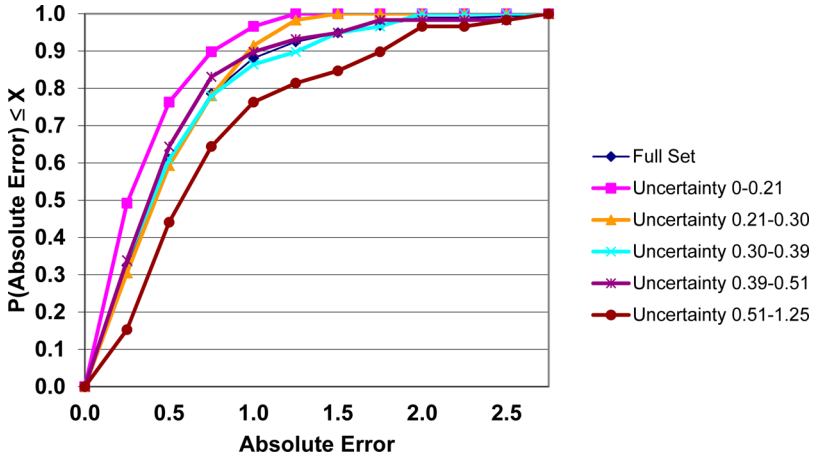


Figure 7. Cumulative probability distributions of absolute prediction errors for 295 test set compounds that have been grouped according to their prediction uncertainties.

strapping procedure can be used to provide meaningful estimates of prediction uncertainties. Examination of trends over all test sets used in this study clearly reveals that compounds with smaller uncertainties are significantly more likely to exhibit smaller absolute prediction errors. The combination of these factors makes Canvas KPLS an attractive method for understanding SAR and optimizing properties of interest based on visualization of the positive/negative atomic contributions to the predictions.

AUTHOR INFORMATION

Corresponding Author

*E-mail: steve.dixon@schrodinger.com.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The collinearity problem in linear regression, the partial least squares pls approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(2) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, 1984.

(3) Hand, D. J.; Yu, K. Idiot's Bayes: Not so stupid after all? . *Int. Stat. Rev.* **2001**, *69*, 385−399.

(4) Dillon, W. R.; Goldstein, M. *Multivariate Analysis, Methods and Applications*; Wiley: New York, 1984.

(5) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.
(6) Klopman, G.; Ptchelintsev, D. Antifungal triazole alcohols: A comparative-analysis of structure−activity, structure teratogenicity and structure therapeutic index relationships using the multiple computer-automated structure evaluation (multi-case) methodology. *J. Comput.-Aid. Mol. Des.* **1993**, *7*, 349−362.
(7) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(8) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(9) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(10) Cramer, D., III; Patterson, D. E.; Bunce, , J .D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(11) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(12) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. Four-dimensional quantitative structure−activity relationship analysis of a series of interphenylene 7-oxabicylcoheptane oxazole thomboxane A2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.
(13) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput. Aid. Mol. Des.* **2006**, *20*, 647−671.
(14) Winkler, D. A.; Burden, F. R. Holographic QSAR of benzodiazepines. *Quant. Struct.-Act. Relat.* **1998**, *17*, 224−231.
(15) *StarDrop*, version 5.2; Optibrium, Ltd.: Cambridge, U.K., 2012.
(16) Cortes, C.; Vapnik, V. Soft Margin Classifier. U.S. Patent 5,650,492, June 17, 1997.
(17) Vapnik, V. Support Vector Method for Function Estimation. U.S. Patent 5,950,146, September 7, 1999.
(18) Farmen, M. W.; Lambert, C. G.; Rusinko, A. R., III; Young, S. S. Statistical Deconvoluting of Mixtures. U.S. Patent 6,434,542, August 13, 2002.
(19) Hurst, J. R.; Heritage, T. W. Molecular Hologram QSAR. U.S. Patent 6,208,942, March 27, 2001.
(20) Rosipal, R.; Trejo, L. J. Kernel partial least squares regression in reproducing Kernel Hilbert Space. *J. Mach. Learn. Res.* **2001**, *2*, 97−123.
(21) Bennet, K. P.; Embrechts, M. J. An Optimization Perspective on Kernel Partial Least Squares. In *Advances in Learning Theory: Methods, Models and Applications*; NATO Science Series III: Computer & Sytems Science; Vol. *190*; Suyken, J. A. K., Horvath, G., Basu, S., Micchelli, C., Vandewalle, J., Eds; IOS Press: Amsterdam, 2003; pp 227−250.
(22) *Canvas*, version 1.5; Schrodinger L.L.C.: New York, 2012.
(23) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−784.
(24) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics* **2012**, *29*, 157−170.
(25) Sheridan, R.; Feuston, R. P.; Maiorov, V. N.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1912−1928.
(26) Nikolova-Jeliazkova, N.; Jaworska, J. An approach to determining applicability domains for QSAR group contribution models: An analysis of SRC KOWWIN. *Altern. Lab. Anim.* **2005**, *33*, 461−470.
(27) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839−849.
(28) Tetko, I. V.; Shushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.
(29) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T., Jr.; Harris, P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): Design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44*, 4339−4358.
(30) D'Alessio, R.; Bargiotti, A.; Metz, S.; Brasca, M. G.; Cameron, A. E.; Marsiglio, A.; Polucci, P.; Roletto, F.; Tibolla, M.; Vazquez, M. L.; Vulpetti, A.; Pevarello, P. Benzodipyrazoles: A new class of potent CDK2 inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1315−1319.
(31) Tao, Z.; Wang, L.; Stewart, K. D.; Chen, Z.; Gu, W.; Bui, M.; Merta, P.; Zhang, H.; Kovar, P.; Johnson, E.; Park, C.; Judge, R.; Rosenbert, S.; Sowin, T.; Lin, N. Structure-based design, synthesis, and biological evaluation of potent and selective macrocyclic checkpoint kinase 1 inhibitors. *J. Med. Chem.* **2007**, *50*, 1514−1527.
(32) Ewing, W. R.; Becker, M. R.; Manetta, V. E.; Davis, R. S.; Pauls, H. W.; Mason, H.; Choi-Sledeski, Y. M.; Green, D.; Cha, D.; Spada, A. P.; Cheney, D. L.; Mason, J. S.; Maignan, S.; Guilloteau, J.; Brown, K.; Colussi, D.; Bentley, R.; Bostwick, J.; Kasiewski, C. J.; Morgan, S. R.; Leadley, R. J.; Dunwiddie, C. T.; Perrone, M. H.; Chu, V. Design and structure−activity relationships of potent and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **1999**, *42*, 3557−3571.
(33) Fevg, J. M.; Pinto, D. J.; Han, Q.; Quan, M. L.; Pruitt, J. R.; Jacobson, I. C.; Galemmo, R. A., Jr.; Wang, S.; Orwat, M. J.; Bostrom, L. L.; Knabb, R. M.; Wong, P. C.; Lam, P. Y. S.; Wexler, R. R. Synthesis and SAR of benzamide factor Xa inhibitors containing a vinically-substituted heterocyclic core. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 641−645.
(34) Brough, P. A.; Aherne, W.; Barril, X.; Borgognoni, J.; Boxall, K.; Cansfield, J. E.; Cheung, K. J.; Collins, I.; Davies, N. G. M.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Finch, H.; Fink, A.; Hayes, A.; Howes, R.; Hubbard, R. E.; James, K.; Jordan, A. M.; Lockie, A.; Martins, V.; Massey, A.; Matthews, T. P.; McDonald, E.; Northfield, C. J.; Pearl, L. H.; Prodromou, C.; Ray, S.; Raynaud, F. I.; Roughley, S. D.; Sharp, S. Y.; Surgenor, A.; Walmsley, D. L.; Webb, P.; Wood, M.; Workman, P.; Wright, L. 4,5-Diarylisoxazole Hsp90 chaperone inhibitors: Potential therapeutic agents for the treatment of cancer. *J. Med. Chem.* **2008**, *51*, 196−218.
(35) Wrobel, J.; Steffan, R.; Bowen, S. M.; Magolda, R.; Matelan, E.; Unwalla, R.; Basso, M.; Clerin, V.; Gardell, S. J.; Nambi, P.; Quinet, E.; Reminick, J. I.; Vlasuk, G. P.; Wang, S.; Feingold, I.; Huselton, C.; Bonn, T.; Farnegardh, M.; Hansson, T.; Nilsson, A. G.; Wilhelmsson, A.; Zamaratski, E.; Evans, M. j. Indazole-based liver X receptor (LXR) modulators with maintained atherosclerotic lesion reduction activity but diminished stimulation of hepatic triglyceride synthesis. *J. Med. Chem.* **2008**, *51*, 7161−7168.
(36) Sheppard, G. S.; Wang, J.; Kawai, M.; Fidanze, S. D.; BaMaung, N. Y.; Erickson, S. A.; Barnes, D. M.; Tedrow, J. S.; Kolaczkowski, L.; Vasudevan, A.; Park, D. C.; Wang, G. T.; Sanders, W. J.; Matei, R. A.; Palazzo, F.; Tucker-Garcia, L.; Lou, P.; Zhang, Q.; Park, C. H.; Kim, K. H.; Petros, A.; Olejniczak, E.; Nettesheim, D.; Hajduk, P.; Henkin, J.; Lesniewski, R.; Davidsen, S. K.; Bell, R. L. Discovery and optimization of anthranilic acid sulfonamides as inhibitors of methionine amino-peptidase-2: A structural basis for the reduction of albumin binding. *J. Med. Chem.* **2006**, *49*, 3832−3849.

(37) Marino, J. P., Jr.; Fisher, P. W.; Hofmann, G. A.; Kirkpatrick, R. B.; Janson, C. A.; Johnson, R. K.; Ma, C.; Mattern, M.; Meek, T. D.; Ryan, M. D.; Schulz, C.; Smith, W. W.; Tew, D. G.; Tomazek, T. A., Jr.; Veber, D. F.; Xiong, W. C.; Yamamoto, Y.; Yamashita, K.; Yang, G.; Thompson, S. K. Highly potent inhibitors of methionine aminopeptidase-2 Based on a 1,2,4-triazole pharmacophore. *J. Med. Chem.* **2007**, *50*, 3777−3785.

(38) Rittle, K. E.; Barrow, J. C.; Cutrona, K. J.; Glass, K. L.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Mirrissette, M. M.; Nantermet, P. G.; Newton, C. L.; Sanders, W. M.; Yan, Y.; Vacca, J. P.; Selnick, H. G. Unexpected enhancement of thrombin inhibitor potency with *o*-ainoalkylbenzylamides in the P1 position. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3477−3482.

(39) An, Y.; Sherman, W.; Dixon, S. L. Hole filling and library optimization: Application to commercially available fragment libraries. *Bioorgan. Med. Chem.* **2012**, *20*, 5379−5387.

(40) Dixon, S. L.; Villar, H. O. Investigation of classification methods for the prediction of activity in diverse chemical libraries. *J. Comput.-Aid. Mol. Des* **1999**, *13*, 533−545.