articles

molecular pharmaceutics

In Silico Drug Profiling of the Human Kinome Based on a Molecular Marker for Cross Reactivity

Xi Zhang[†] and Ariel Fernández*[,†,‡]

Division of Applied Physics and Rice Quantum Institute and Department of Bioengineering, Rice University, Houston, Texas 77005

Received January 20, 2008; Revised Manuscript Received April 15, 2008; Accepted April 30, 2008

Abstract: Protein kinases are paradigmatic targets in molecular cancer therapy. Affinity profiles of kinase drug inhibitors are of considerable interest to assess and modulate clinical drug impact. In the initial stages of discovery, a thorough experimental screening of lead compound libraries becomes critically limited by the size of the kinase sample. This work introduces a computational screening approach which provides a tool for extensive screening that uses experimentally obtained small-scale profiles as input data and makes predictions for a larger kinase set. These predictions result from a propagation of the reduced profile, exploiting a structural comparison of kinases based on a feature-similarity matrix. The comparison focuses on a molecular marker for specificity and promiscuity of kinase inhibitors. Our approach enables the computational high-throughput screening of entire libraries of compounds to search for suitable leads, mapping their inhibitory impact on a sizable sample of the human kinome. Our *in silico* tool is validated by contrasting predictions against reported high-throughput screening experiments.

Keywords: Human kinase; cancer therapy; lead compound; affinity profile; computational screening

Introduction

A primary problem in drug development, to identify compounds with controlled specificity against clinically relevant targets, is usually handled in early stages of development by high-throughput screening, both experimentally and computationally (*in silico*).[1,2] Due to the increasing cost resulting from high clinical failure rates in a downstream stage of development, drug designers are prone to terminate the efforts on those compounds likely to fail in late stages as early as possible.[2] This "fail early" strategy places a significant responsibility on the early stage compound profiling.[3] The two basic types of screening, experimental and computational, are complementary approaches: the former is relatively accurate but cost-limited by the size of the compound library to be profiled,[4] while the latter is more time- and cost-efficient but less reliable.[2,5,6] *In silico* screening methods[7] can be categorized as either ligand-based[8] or target-based.[9] The latter is becoming mainstream for cases

* Corresponding author. Mailing address: Rice University, Department of Bioengineering, 6100 Main St MS142, Houston, TX 77005. Phone: 713-348-3681. Fax: 713-348-3699. E-mail: arifer@rice.edu.

† Division of Applied Physics and Rice Quantum Institute.

‡ Department of Bioengineering.

(1) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960–1964.

(2) Bleicher, K. H.; Böhm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev.* **2003**, *2*, 369–378.

(3) Liszewski, K. Drug Discovery: Successful Lead Optimization Strategies. *Genet. Eng. Biotechonol. News* **2006**, *26*, 14.

(4) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lelias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev.* **2004**, *3*, 935–949.

(6) Shoichet, B. B. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(7) Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349–358.

10.1021/mp800010p CCC: $40.75
Published on Web 07/12/2008

where target structural information is available.[5,7,10] Most of the target-based virtual screening is performed by docking and scoring.[9] However, the inaccuracies of scoring functions pose a major problem in target-based virtual screening.[5,6] Such docking-based algorithms are inadequate to examine kinase targeting[11–19] because most of them do not take into account the induced fits upon binding, which are crucial for kinase−ligand associations that typically involve loopy regions.[11] There exist some docking-based algorithms that take into account induced fits, but they still cannot handle extensive induced fit adaptation including large movements of the backbone,[12,13] which is just the case for the loopy regions in kinases. In fact, kinase ATP-pockets are partly framed by floppy regions, including the activation loop, catalytic loop and P-loop.[11] These parts of the structure undergo an order-upon-binding transition upon association with the ligands (natural or otherwise) which cannot be captured with a docking algorithm, no matter how well it is able to handle flexibility. The induced fit problem as it stands today is almost as hard as the protein folding problem, and no algorithm deals with it effectively from first principles.

This problem notwithstanding, the recent discovery of novel molecular markers for specificity and promiscuity[14] may herald the advent of novel predictive tools for drug affinity profiling by providing a new vantage point for kinase comparisons, as described in this contribution. These comparisons will enable us to predict cross reactivities. Unexpected cross reactivities became recently apparent with the advent of high-throughput experimental screening techniques[4] based on bacteriophage kinase display. Thus, the affinity profiles of 20 inhibitors against a battery of 119 kinases have been reported.[4] However, the operational value of these assays to identify leads from within a compound library (∼1000 compounds) is limited by cost, justifying our development of *in silico* profiling tools.

In this work we introduce a predictive profiler based on the assumption that a structure-based feature-similarity comparison of molecular targets can be used as a surrogate for the differences in their pharmacological behavior.[14] Our approach exploits the recent discovery of a structural marker governing specificity, a result holding even for kinases lacking PDB representation,[14] and, based on this feature, it introduces a comparison of kinases including purported targets and experimentally confirmed targets. The core of our affinity-profile predictor involves determining a linear propagator of profiling data. This propagator consists of the structure-based estimation of pharmacological distances across kinases. Once the propagator is computed, the inference of affinity profiles for test drugs becomes a problem in distance geometry.

Methods

Operational Premises of the Predictor. The identified specificity marker is the solvent-accessible hydrogen bond (SAHB) in the target, inferred from protein structure by quantifying the extent of intramolecular dehydration of the amide−carbonyl pair. The latter parameter gives the number of the side-chain nonpolar groups within a microenvironment around the hydrogen bond, describing its level of solvent exposure (Figure 1). Thus, SAHBs are insufficiently dehydrated and become molecular markers for specificity due to the following properties: (1) SAHBs are sticky since they promote the removal of surrounding water upon associations (Figure 2) and (2) SAHBs are *not conserved across paralog kinases, regardless of their degree of structural similarity* (Figure 3), thus promoting selectivity when and if appropriately targeted.[14] Based on these properties, we introduced the environmental or wrapping distance matrix ($\mathbf{D}_{env}$) for all kinase pairs obtained from the respective differences in the SAHB patterns within aligned structures. On the other hand, the differences in the affinity profiles of kinases against a background of drugs are quantified and arrayed in the so-called pharmacological distance matrix $\mathbf{D}_{phar}$. Our previous work revealed a tight linear correlation between the environmental and the pharmacological distances, thus delineating the molecular basis for cross reactivity. The profiling method presented in this work is an application of this correlation to estimate $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$ and, further, to construct a full affinity profile for a test compound from its subprofile obtained from experimental affinity assays against a small kinase subset. This problem becomes a linear algebra problem, where the full affinity vector for a given drug may

(8) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27–34.
(9) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.
(10) Kuhn, P.; Wilson, K.; Patch, M. G.; Stevens, R. C. The genesis of high-throughput structure-based drug discovery using protein crystallography. *Curr. Opin. Chem. Biol.* **2002**, *6*, 704–710.
(11) Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **2002**, *109*, 275–282.
(12) Mizutani, M. Y.; Itai, A. Efficient Method for High-Throughput Virtual Screening Based on Flexible Docking: Discovery of Novel Acetylcholinesterase Inhibitors. *J. Med. Chem.* **2004**, *47*, 4818–4828.
(13) Mizutani, M. Y.; Takamatsu, Y.; Ichinose, T.; Nakamura, K.; Itai, A. Effective Handling of Induced-Fit Motion in Flexible Docking. *Protein: Struct., Funct., Bioinf.* **2006**, *63*, 878–891.
(14) Chen, J.; Zhang, X.; Fernández, A. Molecular basis for promiscuity and specificity in the druggable kinome. *Bioinformatics* **2007**, *23*, 563–572.
(15) Bain, J.; McLauchlan, H.; Eliott, M.; Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **2003**, *371*, 199–204.
(16) Druker, B. J. Molecularly targeted therapy: have the floodgates opened. *Oncologist* **2004**, *9*, 357–360.
(17) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.
(18) Knight, Z. A.; Shokat, K. M. Features of Selective Kinase Inhibitors. *Chem. Biol.* **2005**, *12*, 621–637.
(19) 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.
(20) Fernández, A. Keeping Dry and crossing membranes. *Nat. Biotechnol.* **2004**, *22*, 1081–1084.

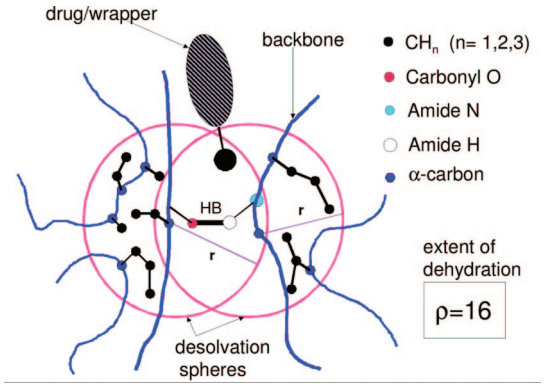


Figure 1. SAHB in a soluble protein. An intramolecular amide−carbonyl hydrogen bond in a soluble protein target prevails provided it is dehydrated (protected from water attack). The extent of dehydration is determined in relation to the bond's microenvironment, defined as two desolvation spheres centered at the alpha-carbons of the hydrogen-bonded residues. Thus, the extent of intramolecular dehydration, ρ, is quantified by the number of side-chain nonpolar groups (black balls) residing within the bond microenvironment.

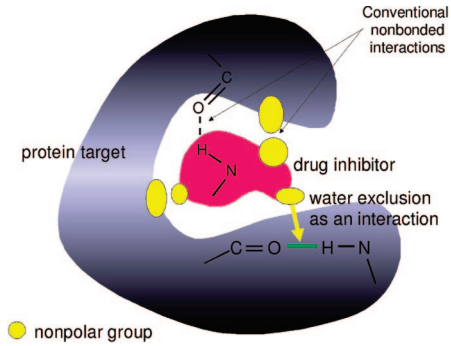


Figure 2. Solvent-accessible hydrogen bond (SAHB) as a targetable binding site in drug−protein interaction. The SAHB is shown as a green segment representing a poorly wrapped amide−carbonyl hydrogen bond in the protein. It promotes the removal of surrounding water upon association. The exclusion of solvating water from pre-existing hydrogen bonds may promote specificity since the SAHB is typically not conserved across paralogs of the protein target.[14] On the other hand, intermolecular pairwise interactions (hydrogen bonds, hydrophobic pairing or charge matching) between protein target and drug ligand typically promote promiscuity due to the high level of amino acid conservation at the ATP-binding region.[14] Shown in the figure are intermolecular hydrogen bond (dashed line) and hydrophobic interaction between nonpolar groups, conventional designers' choices that promote promiscuity.

be determined from suitably defined distances between kinases and the experimentally obtained affinity vector.

The implementation of the profiling method is thus carried out according to the following steps (sketched in Figure 4):

1. Estimate the pharmacological distance matrix ($\mathbf{D}_{phar}$) defined over all kinase pairs by appropriately rescaling the environmental distance matrix ($\mathbf{D}_{env}$). This estimation will

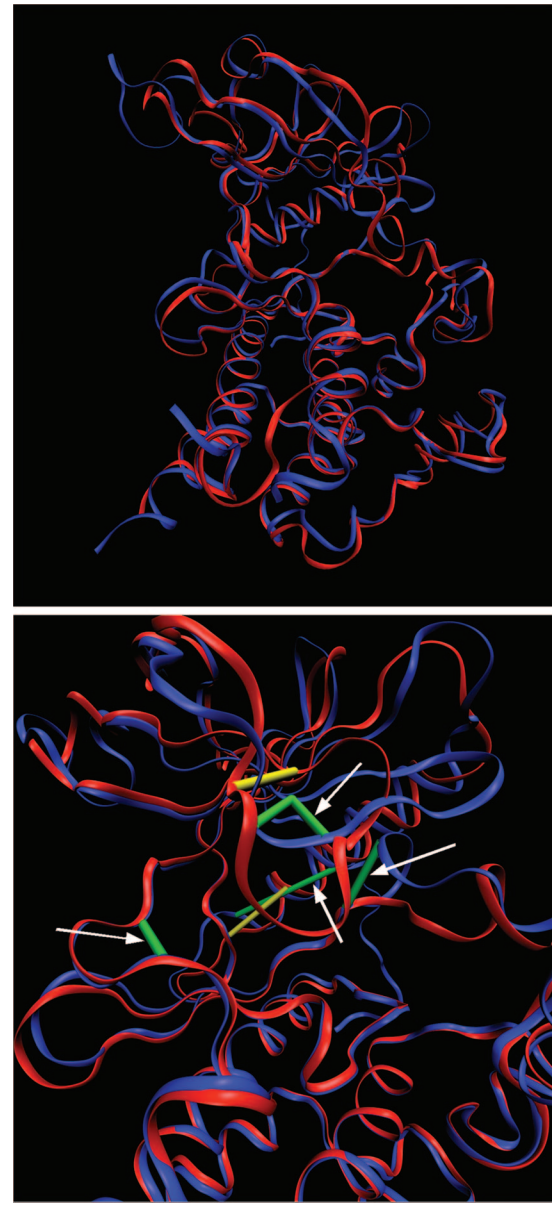


Figure 3. Structural alignment of C-Kit kinase and Bcr-Abl kinase and their SAHB patterns in ATP-binding pockets. The figure highlights the relevance of SAHBs as selectivity filters distinguishing drug targets. (a, top) Structural alignment of C-Kit kinase (pdb code 1T46, ribbon representation in blue) and Bcr-Abl kinase (pdb code 1FPU, ribbon representation in red). The two structures are aligned by DaliLite (http://www.ebi.ac.uk/DaliLite/) and rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). Only backbones are indicated for clarity. The structure similarity (rmsd $\sim$ 1.4 Å) between the two kinases makes it difficult to specifically target one without touching the other. (b, bottom) SAHB distributions in the ATP-binding pockets of C-Kit and Bcr-Abl shown in the same alignment of (a). SAHBs of C-Kit are represented by green virtual bonds joining α-carbons while SAHBs of Bcr-Abl by yellow. While the backbone structures are highly alignable, pointing to high levels of structure conservation as shown in (a), SAHBs are not conserved. The white arrows indicate the nonconserved SAHBs.

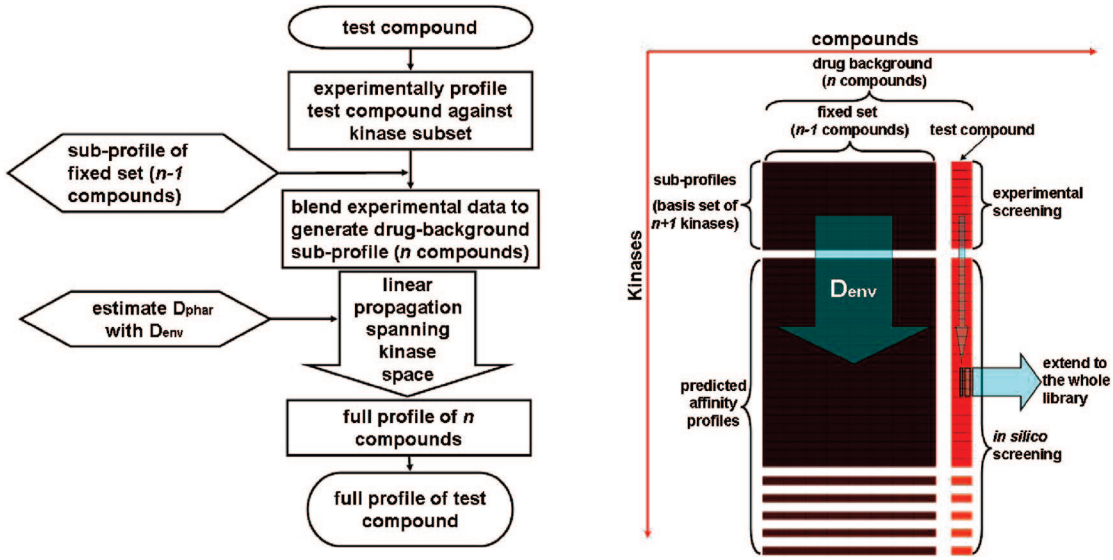


Figure 4. Flowchart (left) and process diagram (right) of the *in silico* profiling method. In the diagram, each column corresponds to one compound and each row to one kinase. The brown columns correspond to the compounds with profiles already known, and the red column corresponds to the test compound. The upper rows represent the subprofiles that are obtained from experiments, and the lower rows represent the profiles predicted by the profiler.

increase in accuracy as a more comprehensive background of drugs, i.e. one covering all kinase targets, is used to determine pharmacological behavior.[14]

2. For every test compound to be fully profiled *in silico*, choose a subset of kinases to obtain a small-scale experimental affinity profile, heretofore noted the "subprofile". The required size of the kinase subset sampled to produce the subprofile depends on the fixed number of inhibitors that define the pharmacological matrix $\mathbf{D}_{phar}$.

3. Determine the full affinity profile for a test compound from its subprofile and our estimated $\mathbf{D}_{phar}$.

4. Repeat the steps above for each test compound in the library.

The subprofile is obtained from small-scale experiments, most advantageous in the case when a battery of a large number of kinases needs to be screened.

Estimating Pharmacological Distances from Environmental Distances. To infer $\mathbf{D}_{phar}$, it becomes necessary to identify the structural feature that governs drug cross reactivity across paralog kinases. Recent research[14] revealed that SAHBs constitute such structural markers. Thus, we estimate $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$ obtained by comparing the SAHB patterns of purported targets. The matrix $\mathbf{D}_{env}$ quantifies differences in the SAHB patterns within the ATP pockets (i.e., the drug binding site) across all kinase pairs. Thus, the environmental distance is based on structural alignment followed by comparison of poorly conserved features.

To determine SAHBs, we assess the hydrogen-bond microenvironment by calculating the extent of intramolecular dehydration, ρ, of the hydrogen bond. In turn, ρ is quantified as the number of the side chain nonpolar groups within a dehydration domain (Figure 1). This domain consists of two intersecting spheres of radius 6.0 Å (~width of three solvation layers[21]) centered at the α-carbons of the hydrogen-bond paired residues. In soluble protein domains, at least two-thirds of the backbone hydrogen bonds lie in the range $\rho = 26.6 \pm 7.5$. SAHBs are thus defined as the hydrogen bonds with ρ equal to or less than 19, i.e. hydrogen bonds lying in the tail of the distribution. SAHBs may be identified using the web-downloadable program YAPView (http://protlib.uchicago.edu/dloads.html). To calculate the SAHBs of a protein, load the pdb file of this protein in YAPView, and choose the proper parameters and options (set radius of desolvation sphere to be 6.0 Å, enable desolvation calculations). The SAHBs are marked as green bars in the protein structure by YAPView.

To compare the SAHB patterns for different kinases, we define the environmental hull of a kinase, H_{env}, as the set of all residues contributing to the microenvironment of hydrogen bonds, i.e. the environmental residues, as well as residues aligning with environmental residues from other paralog chains. The $i-j$ matrix element of $\mathbf{D}_{env}$, associated with kinases i and j, is then defined by comparing the aligned hydrogen-bond microenvironments within the environmental hulls of kinases i and j:

$$D_{env}(i,j) = M(i,j)^{-1}\left[\sum_{n=1,\ldots,M(i,j)} \Delta_n(i,j)\right]$$

where $M(i,j)$ = number of residue pairs in $H_{env}(i)$ corresponding to SAHBs in kinase i or to hydrogen bonds or nonbonded residue pairs that align with SAHBs in $H_{env}(j)$; n = dummy index denoting residue pair; and $\Delta_n(i,j) = 1$ if residue pair n corresponds to a SAHB in $H_{env}(i)$ that aligns with a non-SAHB in $H_{env}(j)$ or *vice versa*, and $\Delta_n(i,j) = 0$,

(21) Fernández, A.; Berry, R. S. Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13460–13465.

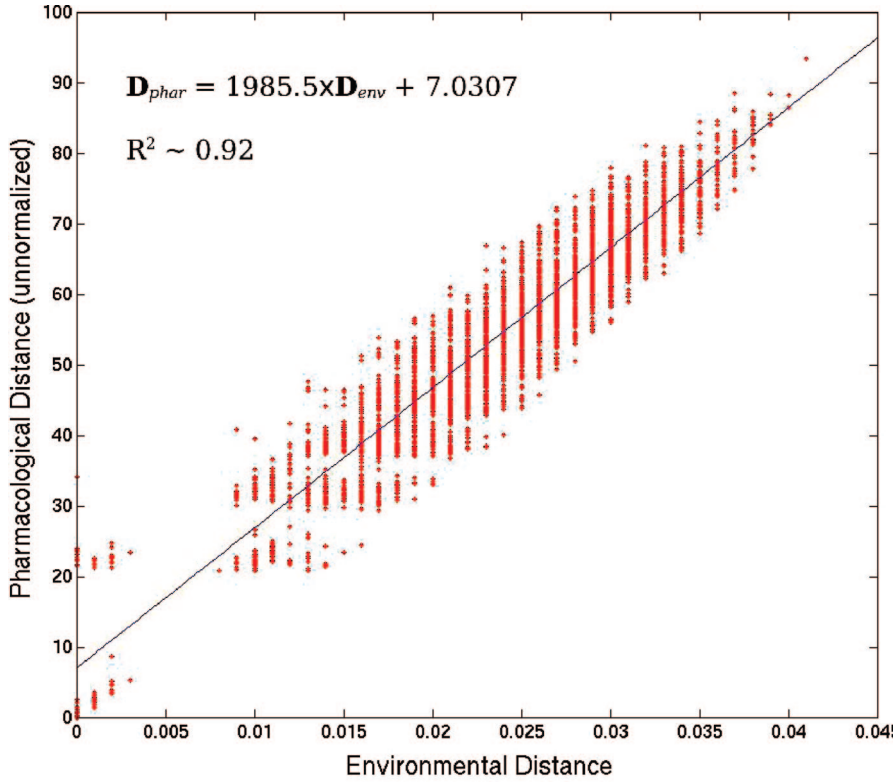


Figure 5. Correlation between environmental and pharmacological distances. Each diamond represents a pair of kinases with horizontal coordinate being the environmental distance and vertical coordinate being the pharmacological distance between them. Unlike Figure 2d in ref 14, the environmental distances in this work are not normalized. The straight line indicates the linear fit by least-squares method: $D_{phar}(i,j) = 1985.5D_{env}(i,j) + 7.0307$. $R^2 \sim 0.92$.

if n corresponds to SAHBs or non-SAHBs in both $H_{env}(i)$ and $H_{env}(j)$. Thus defined, the environmental distance compares local dehydration propensities associated with SAHB patterns in kinases.

On the other hand, $\mathbf{D}_{phar}$ is used to quantify differences in the affinity profiling of kinases against a background of inhibitors.[14,22] $\mathbf{D}_{phar}$ is defined as the matrix of Euclidean distances between affinity vectors for kinase pairs with entries given in negative logarithmic or, equivalently, $\Delta G/RT$-units:

$$D_{phar}(i,j) = \|A_i - A_j\| = \left[\sum_n (A_{i,n} - A_{j,n})^2\right]^{1/2}$$

where A_i and A_j are the affinity profiles of kinases i and j, with components $A_{i,n}$ and $A_{j,n}$ being the values of the negative logarithms of binding constants for complexation of inhibitor n with kinases i and j, respectively. A "no hit" entry in the screening table of Fabian et al.[4] is assigned the value -23.026 $\Delta G/RT$-units, corresponding to a large dissociation constant $K_d = 10^{10}$ μM. Notice that our "no hit" value is well above the cutoff ($K_d = 10$ μM) adopted previously.[14]

Figure 5 shows the correlation between $\mathbf{D}_{phar}$ and $\mathbf{D}_{env}$ when the former is obtained from the T7-bacteriophage kinase display screening against a background of 17 drugs.[4] For each kinase pair (i, j), we can then infer the pharmacological distance using the linearly fitted parameters:

$$D_{phar}(i,j) = 1985.5D_{env}(i,j) + 7.0307$$

Note that the correlation is very tight ($R^2 \sim 0.92$).

While inferring $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$, some errors are introduced in the estimated pharmacological distances. There are at least two possible sources of errors leading to dispersion in the correlation: (a) the background of drugs used to define the pharmacological distance is limited, with uneven target coverage, and thus only approximately indicative of pharmacological behavior; or (b) the SAHBs are not the only selectivity determinant for nonpromiscuous drugs. In regards to error source (a), we are limited in our analysis by the availability of drugs chosen to define pharmacological profiles in high-throughput experiments.[4] We may need to revise $\mathbf{D}_{phar}$ as new screening data becomes available. In regard to (b), we can only claim that SAHBs are a determinant but not necessarily the only factor governing ligand specificity.[14] Even though these errors will be inherited in the following steps, we made the method less sensitive to systematic errors through adequate parametrization.

Expanding Pharmacological Information from Limited Affinity Profiles. We now show how to determine the affinity profiles of kinases from structure-based estimations of pharmacological distances between kinase pairs. Just like vector coordinates cannot be uniquely determined from vector distances, affinity profiles cannot be uniquely determined solely from the pharmacological distances: additional constraints are required. To solve the profile prediction

(22) Fernández, A.; Maddipati, S. A priori inference of cross reactivity for drug-targeted kinases. *J. Med. Chem.* **2006**, *49*, 3092–3100.

problem, we cast it in terms of linear algebra, through a (vector) ↔ (kinase profile) correspondence. The procedure boils down to determining vector coordinates given distances between vectors. To guarantee the uniqueness of the solution, a subset of vectors should be given in addition to vector distances. Thus, for given pharmacological distances between pairs of kinases, the number of independent degrees of freedom of the solution vectors (kinase profiles) is n, which corresponds to the number of inhibitors, i.e. the dimension of the affinity-profile space. Therefore, we need to know at least n independent vectors to determine all affinity profiles. These conditions narrow down solutions to two possibilities, due to symmetry. Reflection relative to the hyperplane determined by the fixed n independent vectors produces two conjugate solutions to the problem. Thus, to unambiguously determine the solution, we need the coordinates of an additional vector. Accordingly, the minimal number of fixed vectors should be $n+1$. A constraint on these $n+1$ vectors is that n of them should be linearly independent.

To summarize, we may recast the profile prediction problem in linear-algebra terms through the following correspondences:

space dimension ↔ number of sampled inhibitors

distance ↔ pharmacological distance

vector ↔ affinity profile of kinase (against a background of inhibitors)

vector $\mathbf{x}_i$ ↔ experimentally determined profile of the ith kinase in the subset

vector $\mathbf{y}$ ↔ affinity profile of the test kinase

Thus, the *in silico* profiling problem may be cast as a linear-algebra problem as follows: In an n-dimension space, we have $n+1$ given vectors (the subset) with n of them being linearly independent:

$$\mathbf{x}_i,\ i = 0, 1, 2,..., n \tag{1}$$

Note that $\mathbf{x}_i$ is the profile vector of kinase i against the n inhibitors. For a generic vector $\mathbf{y}$ (affinity profile of a test kinase) that is not in the subset, we have estimated the distances d_i from $\mathbf{y}$ to all $\mathbf{x}_i$'s:

$$d_i = |\mathbf{y} - \mathbf{x}_i|,\ i = 0,..., n$$

Our purpose is to determine the coordinates of vector $\mathbf{y}$ based on the conditions given above. Note that $\mathbf{y}$ is the profile vector of a kinase that is not in the subset to be screened by experiment, i.e., $\mathbf{y}$ is one of the profile vectors to be determined *in silico*. We now show that this can be achieved by linear algebra calculation.

Since n of the $n + 1$ vectors are linearly independent, we can assume without loss of generality that $\mathbf{x}_1,..., \mathbf{x}_n$ are the independent vectors. We first define $\mathbf{x}_i' = \mathbf{x}_i - \mathbf{x}_0$, for $i = 1, 2,..., n$. Note that all $\mathbf{x}_i'$ are linearly independent. Similarly, we define $\mathbf{y}' = \mathbf{y} - \mathbf{x}_0$. Note that now we only need to determine $\mathbf{y}'$ to obtain $\mathbf{y}$. We can determine the scalar product of $\mathbf{y}'$ with each $\mathbf{x}_i'$ using the information on the vector distances and the well-known relation $|\mathbf{y}' - \mathbf{x}_i'|^2 = |\mathbf{y}'|^2 + |\mathbf{x}_i'|^2 - 2\mathbf{y}' \cdot \mathbf{x}_i'$, in turn obtained from the definition of scalar product:

$$\alpha_i = \mathbf{y}' \cdot \mathbf{x}_i' = (|\mathbf{y}'|^2 + |\mathbf{x}_i'|^2 - |\mathbf{y}' - \mathbf{x}_i'|^2)/2 = (|\mathbf{y}'|^2 + |\mathbf{x}_i'|^2 - d_i^2)/2 \tag{2}$$

where the last expression indicates how α_i can be evaluated from the estimated distances. We then have a system of n-variable linear equations by rewriting the relationships above

$$\begin{aligned} \mathbf{x}_1' \cdot \mathbf{y}' &= \alpha_1 \\ \mathbf{x}_2' \cdot \mathbf{y}' &= \alpha_2 \\ &\vdots \\ \mathbf{x}_n' \cdot \mathbf{y}' &= \alpha_n \end{aligned} \tag{3}$$

Thus the pharmacological problem of determining the profiles of test kinases boils down to solving the system of linear eqs 3, to obtain $\mathbf{y}'$ and thereby $\mathbf{y} = \mathbf{y}' + \mathbf{x}_0$.

Note that each test-kinase profile $\mathbf{y}$ is represented as a row in the "predicted affinity profiles" in the scheme shown in Figure 4. This calculation can be repeated for each test kinase not included in the subset of experimentally screened kinases. By calculating the "predicted affinity profiles" row-by-row, we then extend the subprofile to an entire affinity profile over all kinases for which structural information is available (and hence pharmacological distances may be estimated).

Prediction of Affinity Profiles. In order to find the affinity profile of a new inhibitor, we first obtain its affinity subprofile against the kinases in the subset. We treat the inhibitor as one of the linear dimensions of the pharmacological distance space. As shown above, we calculate the affinities of all the inhibitors, including the new one, toward the kinases that are not within the subset. In this way, the entire profile of the new inhibitor is obtained. In an ideal case as a mathematical model, if the estimated $\mathbf{D}_{\text{phar}}$ were highly close to the real values, we could obtain the quantitatively exact profile (i.e., exact values of K_d's or ΔG's). Unfortunately, the correlation of $R^2 \sim 0.92$ is still not tight enough for a quantitative profile prediction. Alternatively, we predict the profile in a qualitative level, i.e. "hit" or "no hit". At this stage, we use thresholds of the ligand−target dissociation constant ($K_{\text{d,threshold}}$) to determine whether an inhibitor hits a specific kinase or not. We used three thresholds: $K_{\text{d,threshold}} = 1\ \mu$M, 10 μM, and 100 μM.

Repeating these steps on all test compounds, a large library profiling can be generated from a small-scale (i.e., subprofile size) experiment. The process is schematically represented in Figure 4.

Finding the Optimal Basis Set. The choice of kinases in the basis set (the subset) is crucial for the predictor's performance. A required condition for the subset is that the affinity vectors corresponding to the kinases span all dimensions of the affinity space. Furthermore, some of the kinases discriminate the inhibitory compounds less effectively than the others, i.e. compounds' affinities to them are more uniform than that to others. If such kinases are included in

the subset, the prediction would be more sensitive to errors in the estimated pharmacological distances. This would yield larger errors in the predicted affinity vectors. We then designed a simple algorithm to examine and optimize a series of subset choices. First, the predicted results are benchmarked against the experimental results[4] and the accuracies (percentages of the correct predictions) are calculated. We compared our predicted results on 17 inhibitors out of the 20 in the screening experiment[4] excluding three promiscuous inhibitors (Staurosporine, EKB-569 and SU11248), against 119 kinases. For the 17 × 119 entries, we counted the number of entries correctly predicted, in the "hit or no hit" level, and calculated the percentage of correct predictions. Using this percentage as scoring function of the subset-choice, we performed optimizations for the basis set and found several sets with correct prediction percentages around 93%. Our algorithm to optimize the basis set is shown in the pseudo code below:

FOR each kinase in the basis set,
replace this kinase with each kinase not in the set;
predict the profile;
benchmark the predictive results and calculate the accuracy;
IF the current percentage is higher than the existing best one;
THEN use this kinase in the place of the original kinase in the set;
END LOOP

Results

Experimental Validation of Affinity Predictions. The affinity prediction is validated by benchmarking the result for all the nonpromiscuous inhibitory compounds in the phage-display kinase assay against the experimental data.[4] The compounds used in the benchmark are SB202190, SB203580, VX-745, BIRB-796, SP600125, Gleevec, Iressa, Tarceva, ZD-6474, CI-1033, GW-2016, Vatalanib, MLN-518, LY-333531, BAY-43−9006, Roscovitine and Flavopiridol. For randomly chosen kinase subsets prior to any optimization, mostly the accuracies of the predictions are around 80−90%. [Low accuracies (around 60%) are present but rare. These low accuracies arise from the choices of the kinase subsets extremely vulnerable to the errors introduced in previous step. That is also why we need to perform optimizations: to avoid such choices of subsets.] For instance, consider the randomly chosen subset of kinases: AAK1, ABL1, CDK2, EGFR, ERBB2, FLT3, GAK, JNK1, KIT, LCK, p38-alpha, PDGFRB, PHKG1, SLK, SRC, STK10, VEGFR2, YES.

Using this subset, the profiler predicts affinities for all 119 kinases independently assayed[4] with 252 false positives and false negatives out of 17 × 119 predictions, for the affinity threshold 10 μM. The corresponding accuracy is 88%. Notice that the 17 × 119 predictions cover the subprofiles for the subset and the accuracy is calculated based on all 17 × 119 predictions. This does not weaken the validation since when predicting for a kinase within the subset, the input distances are obtained independently, i.e. estimated from environmental distances. In other words, the predictor does not discriminate kinases within or outside of the subset. In a real work, it is useless to predict the affinities of an inhibitor for kinases within the subset, but here we include these results for the sake of validation.

The next step is to optimize the subset of kinases to improve prediction accuracy. For the purpose of validation, we perform the optimization with a "leave-one-out" algorithm. That is, to predict the profile of one inhibitor, we apply the optimization algorithm discussed above to the other 16 inhibitors and thereby get an optimized subset of 17 (= 16 + 1) kinases. These 17 kinases plus one more randomly chosen kinase constitute the subset used for the prediction of the inhibitor. In this way, the inhibitor profile to be predicted is left out of the subset optimization process, and thus the optimized choice of the subset is completely independent of the inhibitor left out, though it might not be the best choice since one kinase in the subset is chosen randomly. It is possible that the randomly chosen kinase plus the 17 optimized ones do not constitute a complete subset (dimension < 17), rendering eq 3 not solvable. In this case, we simply replace the randomly chosen kinase with another random one until the corresponding eq 3 becomes solvable. As an example, the optimized subset used to predict the profile of SB202190 is (the first 16 kinases result from optimization) as follows: INSR, PTK6, CSNK1G1, TEK, Aurora2, EPHA3, PHKG2, BIKE, EPHB4, FYN, TNIK, DAPK2, FGFR1, SLK, PRKAA1, ERBB2, YES, p38-beta (randomly chosen).

We predict the complete affinity profiles of the 17 inhibitors one by one, using the profiling information of the 17 inhibitors against the 18-kinase subset. Before predicting for each inhibior, a subset optimization based on the other 16 inhibitors is carried out. That means the subset used in the profile prediction for each one of the 17 inhibitors is not necessarily the same in each case.

The accuracy of our predictor, as revealed by Figure 6, is quantitatively summarized in Table 1. In accord with Figure 6, we use three thresholds for kinase-inhibitor hit/no-hit results: $K_{d,threshold}$ = 1 μM, 10 μM, and 100 μM. The accuracy of the predictions are 91%, 93% and 93%, respectively. When using $K_{d,threshold}$ = 1 μM, the errors are mainly due to false positives, while for $K_{d,threshold}$ = 10 μM and 100 μM they arise mostly from false negatives. Thus we can choose different $K_{d,threshold}$'s for different purposes. If we are concerned with drug specificity and promiscuity in a clinical context, it is reasonable to use $K_{d,threshold}$ = 1 μM since it imposes a more stringent criterion for affinity, suitable for clinical purposes. If binding is the sole concern, a better choice would be $K_{d,threshold}$ 10 μM or 100 μM, since these filters entail higher sensitivity.

Validating the Predicted Affinity Profile of a Redesigned Version of Imatinib. To further validate our predictor, we focus on a recently developed kinase inhibitor, WBZ_4 (Figure 7), a redesigned version of the powerful anticancer drug imatinib (Gleevec) with higher specificity than the parental compound.[23] The prototype compound WBZ_4 does not belong to the drug background used in the high-throughput screening previously adopted as benchmark

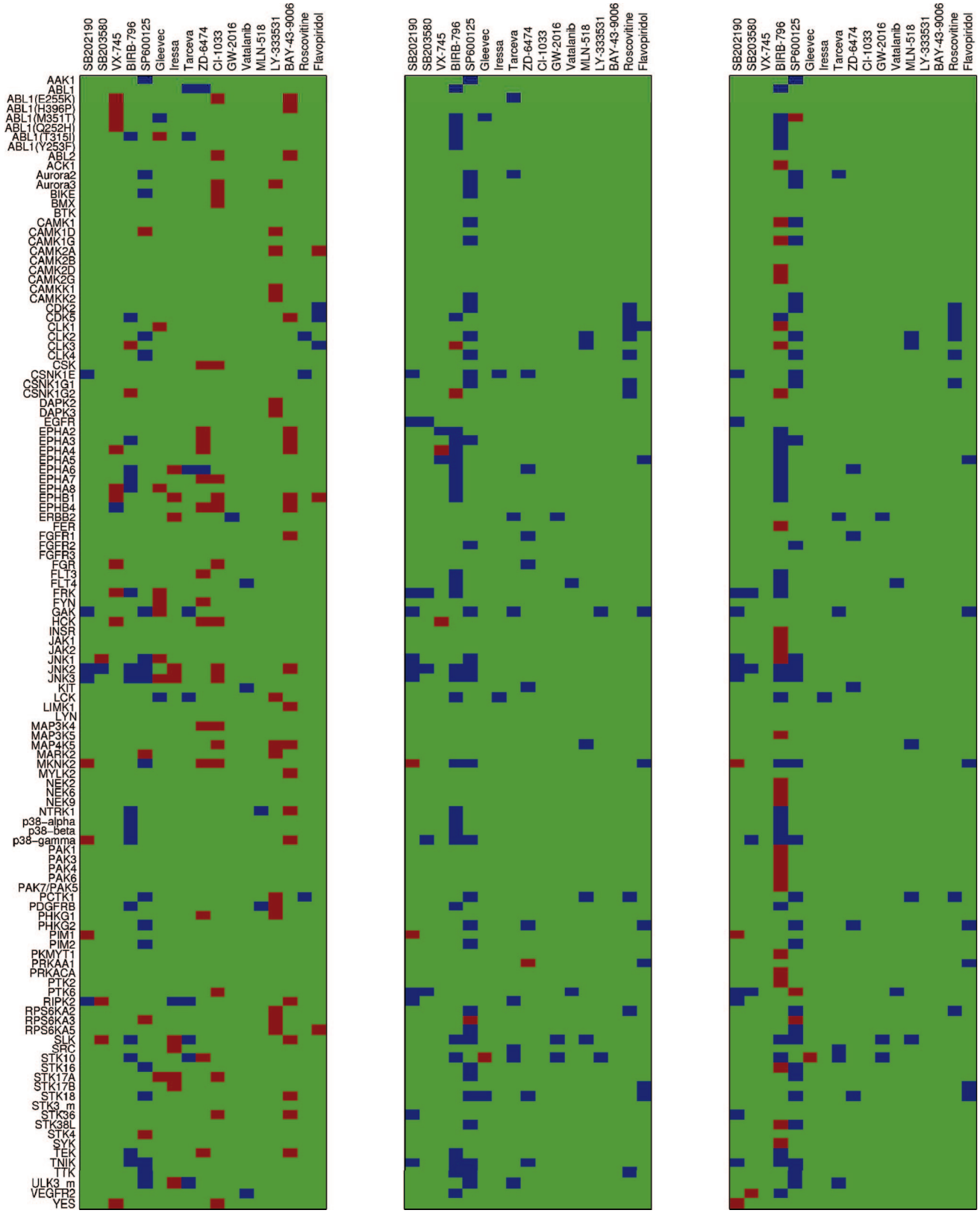


Figure 6. Matrices of prediction performance, corresponding from left to right to affinity thresholds $K_d = 1\ \mu M$, 10 μM, and 100 μM, respectively. The complete affinity profiles of 17 inhibitors independently screened[4] were predicted one by one, using the experimental profiling information on the 17 inhibitors against an 18-kinase subset (different for each inhibitor). Green cells indicate correct predictions; blue, false negatives ("hit" predicted as "no hit"); red, false positives ("no hit" predicted as "hit"). The accuracy percentages shown in the three matrices are 91%, 93% and 93%. Detailed quantitative summary of the accuracy is in Table 1.

Table 1. Prediction Accuracy with Different Filters for Hit/No Hit

$K_{d,threshold}$	1 μM	10 μM	100 μM
false positive[a]	116	9	36
false negative[b]	69	133	114
accuracy	91%	93%	93%

[a] False positive refers to "no hit" predicted as "hit". [b] False negative refers to "hit" predicted as "no hit".

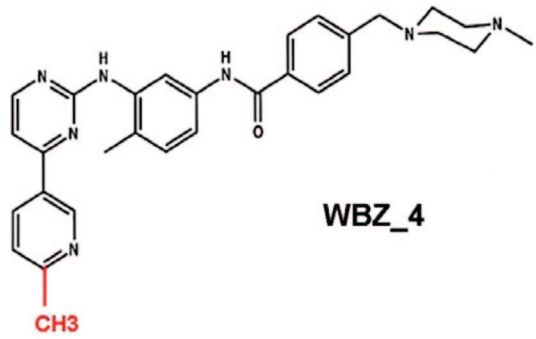


Figure 7. Prototype molecule WBZ_4 (N-{5-[4-(4-methyl piperazine methyl)-benzoylamido]-2-methylphenyl}-4-[3-(4-methyl)-pyridyl]-2-pyrimidine amine). It is developed by adding a methyl group (indicated in red) to the imatinib molecule.

for our method. The interest in this compound arises because the WBZ_4 design was meant to enhance specificity beyond imatinib levels guided precisely by the same structural markers, the SAHBs[24] that we exploited to calculate pharmacological distances and thus infer cross reactivities (Methods). Thus, we now validate our approach by predicting the affinity profile of WBZ_4, and contrasting it with the experimental profile obtained from Ambit's phage-display screening assay reported in ref 23.

The compound WBZ_4 was developed by redesigning imatinib for the purpose of inhibiting the C-Kit kinase, as imatinib does, while avoiding another primary imatinib target, the Bcr-Abl kinase. The latter target has been directly implicated in imatinib's cardiotoxicity.[25] In addition, WBZ_4 was designed to inhibit JNK1, a major target to protect the cardiomyocytes from a mitochondrial collapse induced by Bcr-Abl inhibition.[23,25]

The prototype has been experimentally profiled using the screening methodology based on bacteriophage kinase display.[4,23] In the prediction, the kinase subset has been optimized in consonance with the drug background of 17 nonpromiscuous compounds extracted from the Ambit's screening[4] (Methods). The experimental and the predicted results are contrasted in Figure 8. The experimental results for the affinities of WBZ_4, reported in ref 23, covered 107 of the 119 kinases reported in ref 4, excluding ACK1, Aurora2, Aurora3, NTRK1, PRKAA1, PRKACA, STK10, STK18, STK3_m, STK38L, TEK, and ULK3_m. Due to the emphasis in the pharmacological applications of the prototype compound and the clinical significance of achieving nanomolar inhibition, the theoretical predictions were made adopting a stringent threshold $K_{d,threshold} = 100$ nM, that is, a hit was recorded as such only if $K_d < 100$ nM. Of the 107 predictions, there is no single false negative and there are only 2 false positives: LCK and JNK2, as shown in Figure 8. This corresponds to an accuracy above 98%.

Most importantly, our predictor correctly identified C-KIT and JNK1 as primary targets for WBZ_4 and correctly predicted the lack of pharmacological activity against Bcr-Abl, a crucial premise in the planned imatinib redesign geared at curbing its potential cardiotoxicity.

Comparative Assessment of Performance. One docking-based method was recently reported and claimed to be the best for computing the affinities of inhibitors for homologous receptors.[26] The authors compared their predictions with the experimental results published by Fabian et al.[4] and found "a reasonable but not perfect correspondence".[26] Both this work and our profiler predict the affinity profile of imatinib (Gleevec) against extended lists of kinases, and thus the results can been compared. Benchmarked by the experimental profiles,[4] the docking-based prediction contains 9 false negatives (CLK1, CLK4, EPHA8, GAK, JNK1, JNK2, JNK3, STK17A, STK18) and 8 false positives (ACK1, BMX, CSK, FGR, HCK, LYN, RIPK2, YES) out of the 119 kinases, while our profiler has only 2 false negatives (ABL1(T315I), STK18) and 1 false positive (STK10) with the 10 μM threshold (Figure 6). [The docking-based prediction[26] calculated the affinities of Gleevec for 493 human protein kinases, which are almost the whole human kinome. The 9 false negatives and 8 false positives are only the ones within the 119 kinases tested in the experiments[4].]

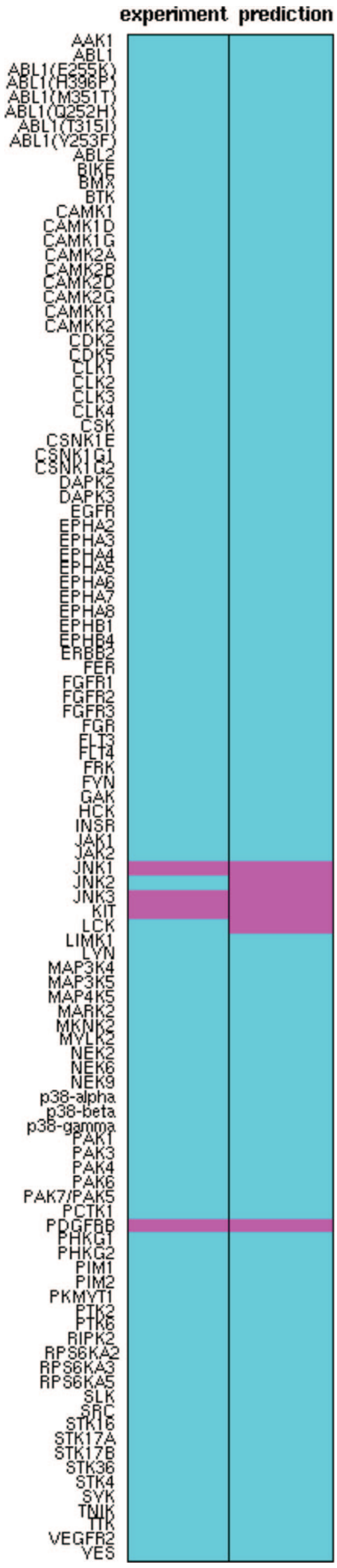
Furthermore, the docking-based method is more demanding than our profiler in terms of primary experimental data, since it requires the "high-resolution structure in complex with at least one protein kinase target" for each inhibitor to make appropriate prediction.[26] The successful generation of such data is plagued with experimental uncertainty as crystallization remains a serendipitous craft rather than an automated methodology. By contrast, our profiler requires the affinity subprofile of the inhibitor against ~20 protein kinases, which can be routinely obtained through phage displayers of kinase batteries or other screening methods.

(23) Fernández, A.; Sanguino, A.; Peng, Z.; Ozturk, E.; Chen, J.; Crespo, A.; Wulf, S.; Shavrin, A.; Qin, C.; Ma, J.; Trent, J.; Lin, Y.; Han, H. D.; Mangala, L. S.; Bankson, J. A.; Gelovani, J.; Samarel, A.; Bornmann, W.; Sood, A. K.; Lopez-Berestein, G. An anticancer C-kit kinase inhibitor is re-engineered to make it more active and less cardiotoxic. *J. Clin. Invest.* **2007**, *117*, 4044–4054.

(24) Crespo, A.; Fernández, A. Kinase packing defects as drug targets. *Drug Discovery Today* **2007**, *12*, 917–923.

(25) Kerkela, R.; Grazette, L.; Yacobi, R.; Iliescu, C.; Patten, R.; Beahm, C.; Walters, B.; Shevtsov, S.; Pesant, S.; Clubb, F. J.; Rosenzweig, A.; Salomon, R. N.; Van Etten, R. A.; Alroy, J.; Durand, J.-B.; Force, T. Cardiotoxicity of the cancer therapeutic agent imatinib mesylate. *Nat. Med.* **2006**, *12*, 908–916.

(26) Rockey, W. M.; Elcock, A. H. Rapid Computational Identification of the Targets of Protein Kinase Inhibitors. *J. Med. Chem.* **2005**, *48*, 4138–4152.



The performance and confidence of our predictor are affected by two factors: error sources and the predictor's vulnerability to errors. The major error source in this algorithm arises from the computational estimation of kinase pharmacological distances from environmental distances, which is determined by the tightness of the correlation. Thus, the predictor may perform not so well in the cases where the profile of the compound to predict is not subject to the correlation. The reasons why the correlation is not perfect and the factors affecting the correlation have already been discussed in previous work and may include alternative structural features nonconserved across paralogs.[14] Here we discuss the practical aspect of the problem: for what kinds of compounds would the predictor work successfully? The 17 compounds we adopted in our analysis represent various types of compounds: SB202190, SB203580, and SP600125 are research compounds; MLN518 is in phase I; VX745, CI-1033, ZD6474, Roscovitine, and Flavopiridol are in phase II; BIRB-796, GW-2016, Vatalanib, LY-333531, and BAY-43-9006 are in phase III; Gleevec, Iressa, and Tarceva are approved drugs.[4] This diversity suggests that the predictor would work well in a wide range of compounds. However, there are highly promiscuous compounds, such as Staurosporine, whose profile cannot be fitted into the structure−pharmacology correlation[14] and hence our predictor would fail to yield a reliable result.

Another important factor influencing predictor confidence is its robustness or vulnerability to errors introduced in the distance estimation. This is mainly determined by the choice of kinases in the small-scale sample, the affinities for which constitute the subprofile. Some of the kinases differentiate the inhibitory compounds better than others, i.e. compounds' affinities for them are more rigorous. It is better to include such kinases in the subprofile, since such basis subsets are less sensitive to the errors in the estimated pharmacological distances. This problem is handled by the optimization

Figure 8. Experimental and predicted results for the affinity profile of WBZ_4 against 107 kinases. The experimental results for the affinities of WBZ_4, reported in ref 23, covered 107 of the 119 kinases reported in ref 4, excluding ACK1, Aurora2, Aurora3, NTRK1, PRKAA1, PRKACA, STK10, STK18, STK3_m, STK38L, TEK, and ULK3_m. The subset adopted in this prediction has been optimized in advance, within a training set excluding WBZ_4. The optimized subset contains ABL1(E255K), CAMK1, EPHA8, ERBB2, FLT3, FRK, GAK, INSR, JNK1, KIT, MAP3K4, PDGFRB, PHKG1, PIM1, PRKAA1, RPS6KA2, SLK, SRC. Due to the emphasis in the pharmacological applications of the prototype compound and the clinical significance of achieving nanomolar inhibition, the theoretical predictions were made adopting a stringent threshold $K_{d,threshold} = 100$ nM, that is, a hit was recorded as such only if $K_d < 100$ nM. Of the 107 predictions, there is no single false negative and there are only 2 false positives: LCK and JNK2. This corresponds to an accuracy above 98%.

process, in which the subset of kinases that performs best within the training set is chosen.

Discussion

In this work, we introduced a method to predict affinity profiles of inhibitor compounds against entire batteries of human kinases based on a structural descriptor of the targets. The method is rooted in a molecular marker governing drug specificity and promiscuity established in recent work. A feature-similarity matrix constructed based on the molecular marker is defined across kinase targets and used as an information propagator of a subprofile where drugs are screened against a small group of kinases. The method reported makes use of distance-geometry techniques, and boils down to determining vectors (kinase profiles) from distances between vectors (feature-similarity distances regarded as surrogates for pharmacological distances between kinases). To guarantee the uniqueness of the solution, some vector coordinates need to be fixed and adopted as constraints. These constraints represent the linear algebra counterpart of the subprofile. Our *in silico* method enables us to screen large libraries of compounds predicting their profile. To provide the input data, only a limited experimental screening against a reduced subset of kinases needs to be performed in advance. Thus, our predictor becomes a valuable tool for lead discovery.

The linear-algebra operations subsumed in our predictor are based on the construction of the informational propagator and hence do not entail any source of errors. The only systematic source of uncertainty arises from the estimation of pharmacological parameters from structure-based attributes (environmental distances). Given the high accuracy (for a qualitative prediction) of the pharmacological distance estimation, our profiling method should be deemed highly reliable: the accuracy is up to 93% with an optimized choice of kinase subset as starting point, as shown above.

Alternative *in silico* screening methods rooted in docking algorithms are unlikely to match this level of accuracy, not only because of their inherent parametric uncertainty and time expense but also because kinase binding entails extensive induced fit of the loopy regions within the ATP-pocket.[11] Even those docking algorithms that incorporate induced fits into affinity calculations cannot handle the lengthy loopy regions of kinases, which undergo extensive structural adaptation.[12,13] The induced fit problem as it stands today remains intractable from first-principle aproaches. This is the main reason why we adopted an information-based algorithm for our predictor.

Acknowledgment. This research was supported by NIH Grant R01 GM072614. We thank Dr. Wotao Yin, Dr. Alejandro Crespo, Jianping Chen and Natalia Pietrosemoli for valuable insights.

MP800010P