

# Structure-Guided Selection of Specificity Determining Positions in the Human Kinome

Mark Moll

Department of Computer Science  
Rice University  
Houston, TX, USA  
mmoll@rice.edu

Paul W. Finn

University of Buckingham  
Buckingham, United Kingdom  
paul.finn@buckingham.ac.uk

Lydia E. Kavraki

Department of Computer Science  
Rice University  
Houston, TX, USA  
kavraki@rice.edu

**Abstract**—It is well-known that inhibitors of protein kinases bind with very different selectivity profiles. This is also the case for inhibitors of many other protein families. A better understanding of binding selectivity would enhance the design of drugs that target only a subfamily, thereby minimizing possible side-effects. The increased availability of protein 3D structures has made it possible to study the structural variation within a given protein family. However, not every structural variation is related to binding specificity. We propose a greedy algorithm that computes a subset of residue positions in a multiple sequence alignment such that structural and chemical variation in those positions helps explain known binding affinities. By providing this information, the main purpose of the algorithm is to provide experimentalists with possible insights into how the selectivity profile of certain inhibitors is achieved, which is useful for lead optimization. In addition, the algorithm can also be used to predict binding affinities for structures whose affinity for a given inhibitor is unknown. The algorithm’s performance is demonstrated using an extensive dataset for the human kinome, which includes a large and important set of drug targets. We show that the binding affinity of 38 different kinase inhibitors can be explained with consistently high precision and accuracy using the variation of at most six residue positions in the kinome binding site.

## I. INTRODUCTION

Predicting affinity profiles remains a challenging task for computational and medicinal chemists. This is particularly true of the kinase family of enzymes because of their large number and structural similarity. Despite their structural similarity, the kinases exhibit large phylogenetic diversity. As a result, binding site sequence dissimilarity alone cannot explain the differences in binding affinity [1]. Selectivity patterns obtained by experimental screening in enzyme assays are often difficult to rationalize in structural terms. Additional tools are needed to improve our capabilities to design inhibitors that *selectively* bind to only a small subset of the kinases. The rapidly increasing number of kinase structures has made it possible to study how structural differences affect binding affinity. For instance, different inhibitors have been designed to target the inactive, DFG-out conformation and active, DFG-in conformation [2–5]. In general, determining exactly how functional changes relate to structural ones remains an important open challenge [6, 7]. This is caused in part by the fact that not all structural changes

cause a functional change. Additionally, the available structures are non-uniformly distributed over the known kinase sequences: for many kinases there is no structural information, while other kinases are overrepresented, which can lead to overfitting.

In previous work [1], we introduced the Combinatorial Clustering Of Residue Position Subsets (CCORPS) method and demonstrated that it could be used to predict binding affinity of kinases. CCORPS considers structural and chemical variation among all triplets of binding site residues and identifies patterns that are predictive for some externally provided labeling. The labeling can correspond to, e.g., binding affinity, Enzyme Commission classification, or Gene Ontology terms, and only needs to be defined for *some* of the structures. CCORPS corrects for the non-uniform distribution of structures. From the patterns CCORPS identifies, multiple predictions are combined into a single consensus prediction by training a Support Vector Machine. A limitation of this work is that it is difficult to identify the most important Specificity Determining Positions (SDPs). In this paper, we are not trying to construct a better predictor, but, rather, a better explanation for some labeling. The explanation is better in the sense that it provides a simple explanation of a labeling in terms of the dominant SDPs. Rather than using *all* patterns discovered by CCORPS, it uses a small number of patterns that involve only a small number of residues yet is able to accurately recover binding affinity.

The main contribution of this paper is an algorithm that computes the Specificity Determining Positions that best explain binding affinity in terms of structural and chemical variation. More generally, the algorithm can identify a sparse pattern of structural and chemical variation that corresponds to an externally provided labeling of structures. This work extends our prior work on CCORPS, but shifts the focus from optimal predictions to concise, biologically meaningful, explanations of functional variation.

The rest of the paper is organized as follows. In the next section we discuss related work. In section III we briefly summarize the CCORPS framework, which forms the basis for our work. Our algorithm for computing SDPs is presented in section IV. The algorithm is evaluated on an extensive kinase dataset in section V. Finally, we end with a brief conclusion in section VI.

## II. RELATED WORK

There has been much work on the identification and characterization of functional sites. Most of the techniques are broadly applicable to many protein families, but we will focus in particular on their application to kinases, when possible.

Much of the work on computing SDPs is based on evolutionary conservation in multiple sequence alignments (see, e.g., [8–10]). There has also been work on relating mutations to an externally provided functional classification in a phylogeny-independent way [11, 12]. This work is similar in spirit to what CCORPS does, but based on sequence alone.

While sequence alignment techniques can reveal functionally important residues in kinases [13], structural information can provide additional insights. This is especially true for large, phylogenetically diverse families such as the kinases. The FEATURE framework [14, 15] represents a radically different way of identifying functional sites. Instead of alignment, FEATURE builds up a statistical model of the spatial distribution of physicochemical features around a site.

Another approach to modeling functional sites has been the comparison of binding site cavities [3, 16]. In [17] a functional classification of kinase binding sites is proposed based on a combination of geometric hashing and clustering. This approach is similar in spirit to our prior work [1], but our work considers variations in a small sets of binding site residues, which makes it possible to separate non-functional structural changes from functional ones.

In [18] many of the ideas above are combined into one framework. Given sequences from a PFAM alignment [19] and some reference structures, homology models are constructed for all sequences. Next, cavities are extracted, aligned, and clustered. Unlike our work, the approach in [18] is completely unsupervised and does not aim to provide an explanation for an externally provided classification.

## III. CCORPS OVERVIEW

Our algorithm builds on the existing CCORPS framework [1]. CCORPS is a semi-supervised technique that takes as input a set of partially labeled structures and produces as output the predicted labels for the unlabeled structures. Of course, this is only possible if the labels can be related to variations in the structures. In previous work [1] we have shown this to be the case for labelings based on binding affinity and functional categorization (Enzyme Commission classification).

CCORPS [1] consists of several steps. First, a one-to-one correspondence needs to be established between relevant residues (e.g., binding site residues) among all structures. This correspondence can be computed using a multiple sequence alignment or using sequence independent methods [20–23]. Second, we consider the structural and physicochemical variation among all structures and all triplets of residues. The triplets are not necessarily consecutive in the protein sequence and can be anywhere in the binding site. Each triplet of residues constitutes a *substructure*: a spatial arrangement of residues. For each triplet, we compute a distance matrix of all pairwise distances between substructures. The distance measure used is

a combination of structural distance and chemical dissimilarity introduced in [21]. In particular, the distance between any two substructures  $s_1$  and  $s_2$  is defined as:

$$\begin{aligned} d(s_1, s_2) = & d_{\text{side chain centroid}}(s_1, s_2) + d_{\text{size}}(s_1, s_2) \\ & + d_{\text{aliphaticity}}(s_1, s_2) + d_{\text{aromaticity}}(s_1, s_2) \\ & + d_{\text{hydrophobicity}}(s_1, s_2) + d_{\text{hbond acceptor}}(s_1, s_2) \\ & + d_{\text{hbond donor}}(s_1, s_2). \end{aligned}$$

The  $d_{\text{side chain centroid}}(s_1, s_2)$  term is the least root-mean-square deviation of the pairwise-aligned side chain centroids of the substructures. The remaining terms account for differences in the amino acid properties between the substructures  $s_1$  and  $s_2$  as quantified by the pharmacophore feature dissimilarity matrix as defined in [21].

Each row in the distance matrix can be thought of as a “feature vector” that describes how a structure differs from all others with respect to a particular substructure. The  $n \times n$  distance matrix for  $n$  structures is highly redundant and we have shown that the same information can be preserved in a 2-dimensional embedding computed using Principal Component Analysis [24]. Each 2D point is then a reduced feature vector. The set of  $n$  2-dimensional points is clustered using Gaussian Mixture Models in order to identify patterns of structural variation. Not all structural variation is relevant; we focus on patterns of structural variation that align with the classification provided by the labeling.

The final stage of CCORPS is the prediction of labels for the unlabeled structures. Suppose a cluster for one of the residue triplets contains structures with only one type of label as well as some unlabeled structures. This would suggest that the predicted label for the unlabeled structures should be the same as for the other cluster members. We call such a cluster a *Highly Predictive Cluster* (HPC). These HPCs are a critical component of the algorithm presented in the next section. There are many clusterings and each clustering can contain several HPCs (or none at all). For example, in the human kinome the binding site consists of 27 residues, leading to  $\binom{27}{3} = 2,925$  clusterings. Typically, an unlabeled structure belongs to several HPCs and we thus obtain multiple predictions. These predictions might not agree with each other. In our prior work we trained a Support Vector Machine to obtain the best consensus prediction from the multiple predictions.

## IV. STRUCTURE-GUIDED SELECTION OF SPECIFICITY DETERMINING POSITIONS

While CCORPS has been demonstrated to make accurate predictions, it has been difficult to interpret the structural basis for these predictions. This has motivated us to look at alternative ways to interpret the clusterings produced by CCORPS. Rather than trying to build a better predictor, we have developed an algorithm that constructs a *concise structural explanation* of a labeling. It determines a set of Specificity Determining Positions (SDPs). An algorithm that would predict that almost every residue position is important would not be very helpful. We therefore wish to enforce a sparsity constraint: for a set

of labeled structures  $S$  we want to find the smallest possible number of HPCs that cover the largest possible subset of  $S$  and involve at most  $\lambda$  residues.

The problem of finding SDPs can be formulated as a variant of the set cover problem. The set cover problem is defined as follows: given a set  $S$  and subsets  $S_i \subseteq S, i = 1, \dots, n$ , what is the smallest number of subsets such that their union covers  $S$ ? This is a well-known NP-Complete problem, but the greedy algorithm that iteratively selects the subset that expands coverage the most can efficiently find a solution with an approximation factor of  $\ln|S|$ .

As mentioned above, in our case,  $S$  is the set of *labeled* structures. We keep track of the residues involved in the selected HPCs and mark them as SDPs. Solving this as a set cover problem would likely still select most residues. The intuition for this can be understood as follows. The number of clusterings each residue is involved in is quadratic in the number of residues in the alignment. Each of those clusterings could contain a HPC that covers at least one structure that is not covered yet by other HPCs. Even in completely random data some patterns will appear, which could in turn be classified as HPCs.

We measure sparsity of the cover in terms of the number of residues and not the number of HPCs, since this facilitates an easier interpretation of the results shown later on. As noted before, there can be several HPCs per clustering. This means that once we have selected an HPC, we might as well include all other HPCs from that same clustering (we have already “paid” for using the corresponding residues). As an algorithmic refinement, we may also wish to limit the degree at which we are fitting the data to avoid overfitting and get a simpler description of the *most significant* residues positions whose variation can be used to explain the labeling.

The algorithm for computing SDPs is shown in Algorithm 1. It is similar to the greedy set cover algorithm. The input to the algorithm consists of a list of labeled structures, a list of all 3-residue subsets of the binding site, and a list of sets of structures that belong to HPCs. After initializing the set of SDPs and the set of selected subset indices in  $S$ , the main loop performs the following steps. First, the indices of all subsets are computed that will not grow the set of SDPs beyond a size limit  $\lambda$  (line 5). Second, the subset index is computed that will increase the cover of the known labels with HPC structures the most (line 9). Next, the algorithm checks whether the increase is “large enough,” i.e., greater than or equal to  $\delta$  (line 11). If so, the set of SDPs and the sets of not-yet-covered structures are updated (line 13–14). If not, the algorithm terminates and returns the set of SDPs.

The final output of Algorithm 1 provides a concise explanation of which structural and chemical variations correlate highly with a given labeling. In the context of the kinases, it can identify triplets of residues whose combined structural and chemical variation give rise to patterns that allow one to separate binding from non-binding kinases. As we will see in the next section, often only a very small set of residues is sufficient to obtain HPCs that cover most of the structures with known binding affinity.

---

### Algorithm 1 Compute Specificity Determining Positions

---

**getSDPs**( $L, S, H, \lambda, \delta$ )

**Input:**  $L$ : set of all labeled structures

**Input:**  $S$ : list of all 3-residue subsets of binding site

**Input:**  $H$ : list of sets of labeled structures s.t.  $H_i$  contains the structures that belong to HPCs in the clustering for subset  $S_i$

**Input:**  $\lambda, \delta$ : parameters that control sparsity and overfitting, respectively.

**Output:**  $P$ : a set of SDPs that best explains the labeling

```

1:  $P \leftarrow \emptyset$  // Set of SDPs
2:  $C \leftarrow \emptyset$  // Set of subset indices in  $S$  chosen so far
3: loop
4:   //  $\lambda$  controls sparsity of SDPs
5:    $I \leftarrow \{i \mid i \notin C \wedge |S_i \cup P| \leq \lambda\}$ 
6:   if  $I = \emptyset$  then
7:     break // No more subsets satisfy sparsity constraints
8:   // Greedy selection of next subset
9:    $j \leftarrow \arg \max_{i \in I} |L \cap H_i|$ 
10:   $C \leftarrow C \cup \{j\}$ 
11:  if  $|L \cap H_j| < \delta$  then
12:    break // Not enough improvement possible
13:   $P \leftarrow P \cup S_j$ 
14:   $L \leftarrow L \setminus H_j$ 
15: return  $P$ 

```

---

## V. RESULTS FOR THE HUMAN KINOME

In [25] a quantitative analysis is presented of 317 different kinases and 38 kinase inhibitors. For every combination of a kinase and an inhibitor, the binding affinity was experimentally determined. This dataset also formed the basis for the evaluation of CCORPS [1]. The kinase inhibitors vary widely in their selectivity. Inhibitors like Staurosporine bind to almost every kinase, while others like Lapatnib bind to a very specific subtree in the human kinase dendrogram. The structure dataset was obtained by selecting all structures from the Pkinase and Pkinase\_Tyr PFAM alignments [19]. The binding site, as defined in [1], consists of 27 residues. After filtering out structures that had gaps in the binding site alignment, 1,958 structures remained. The binding affinity values were divided into two categories (i.e., labels): “binds” and “does not bind.” This gives rise to two different types of HPCs: clusters predictive for binding (which we call *true*-HPCs below) and clusters predictive for *not* binding (which we call *false*-HPCs below). All other structures corresponding to kinases that were not part of the Karaman et al. study [25] do not have a label. CCORPS was run on this dataset, consisting of all 1,958 structures along with the binding affinity data. This resulted in  $\binom{27}{3} = 2,925$  clusterings, one for every triplet of residues. The median number of *true*-HPCs per inhibitor was 591, while the median number of *false*-HPCs per inhibitor was 13,632.

In the next subsection we look in detail at results of our algorithm with one parameter setting to get a sense of what kind of output is produced. In subsection V-B we will then

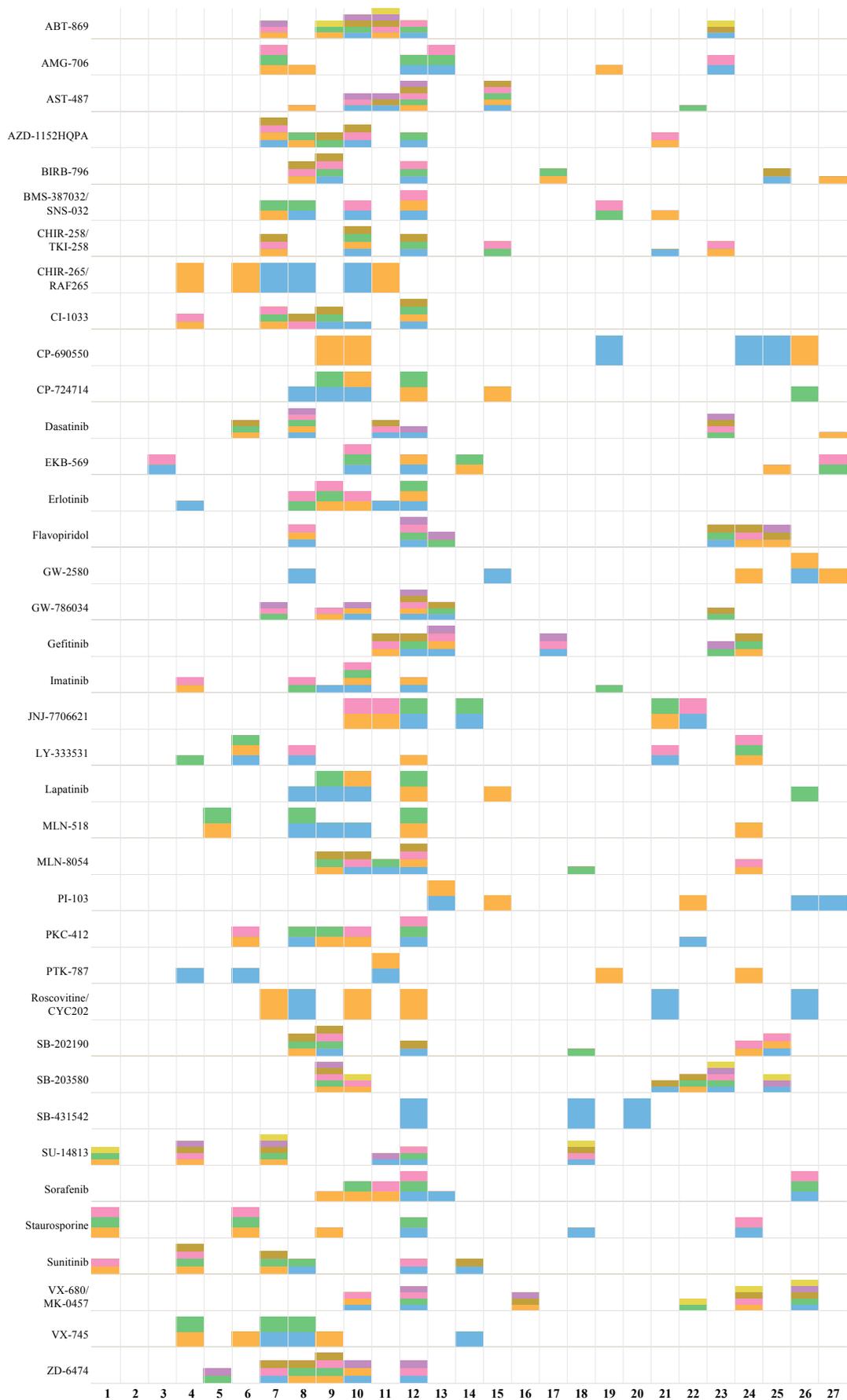


Fig. 1: The SDP profiles computed for every inhibitor in the kinome dataset. The  $x$ -axis represents the residue position in the 27-residue multiple sequence alignment of the binding site. Each row shows the SDPs for one inhibitor whose name is shown on the  $y$ -axis. For each inhibitor, blocks with the same color correspond to one of the 3-residue subsets. If there are multiple colors in a given position, then the same residue was part of several selected subsets. This means that the same residue in different structural contexts can help explain the binding affinity of different kinases.

describe different ways to measure coverage of the SDPs as well as their predictive potential. We then evaluate these measures on all inhibitors with different parameter settings.

### A. Specificity-Determining Positions

While in our prior work [1] the emphasis was on *predicting* the affinity of kinases, here we are focused on creating a *concise explanation* of the affinity. Thus, here we are not performing cross validation experiments. We have run Algorithm 1 on the kinome dataset with  $\lambda = 6$  residues and  $\delta = 16$  (statistics for different values of  $\lambda$  and  $\delta$  are reported in the next subsection). With  $\lambda = 6$ , the algorithm can select at most two non-overlapping triplets. We computed the SDPs for all inhibitors (see Fig. 1). With some additional bookkeeping we can keep track of which residue was involved in which selected subsets. The bar chart for each inhibitor can be interpreted as follows. Along the  $x$ -axis is the residue position in the multiple sequence alignment of the 27 binding site residues. The relative height of each bar indicates how often a residue position was part of a selected 3-residue subset. Blocks with the same color correspond to residues belonging to the same residue subset. This can provide important contextual information. It shows not only *which* residues are important to help explain binding affinity, but also the context in which its variation should be seen. It could, e.g., indicate that one residue’s variation *relative to* some other residue(s) is important. The contextual residues themselves may not always vary much and are perhaps not of as much functional importance in the traditional sense. As  $\lambda$  is increased, more bars would be added to each profile as long as they improve coverage by at least  $\delta$  structures. Similarly, as  $\delta$  is decreased, more bars would be added to each profile as long as no more than  $\lambda$  residues are involved.

Figure 2 shows some examples of the clusterings that have been selected by Algorithm 1. These clusterings contain a large number of structures belonging to HPCs. The distance between points represents how different the corresponding structures are, structurally and chemically. The examples show that we can identify very strong spatial cohesion among the structures that bind when looking at the right residues (i.e., the SDPs). Not all clusterings selected by Algorithm 1 show such a strong relationship between structure and function. Especially for inhibitors that bind more broadly to kinases this relationship is harder to untangle.

There is significant variation among the SDP profiles. For a very selective inhibitor like SB-431542 the variation of only three positions is sufficient to explain the binding affinity (see also the next subsection), while for ABT-869 many combinations of 3 residues out of the 6 selected residues seem to be helpful in explaining the binding affinity.

Fig. 3 shows two different visualizations of the SDPs for the inhibitor Imatinib. Fig. 3(a) shows the structural variation (or lack thereof) in the selected residue positions for all structures that bind Imatinib. Fig. 3(b) shows the sequence logo for those same residues and structures as created by WebLogo [26]. In comparison with Fig. 3(c), we see that the SDPs are much more conserved. Sequence conservation alone is typically not

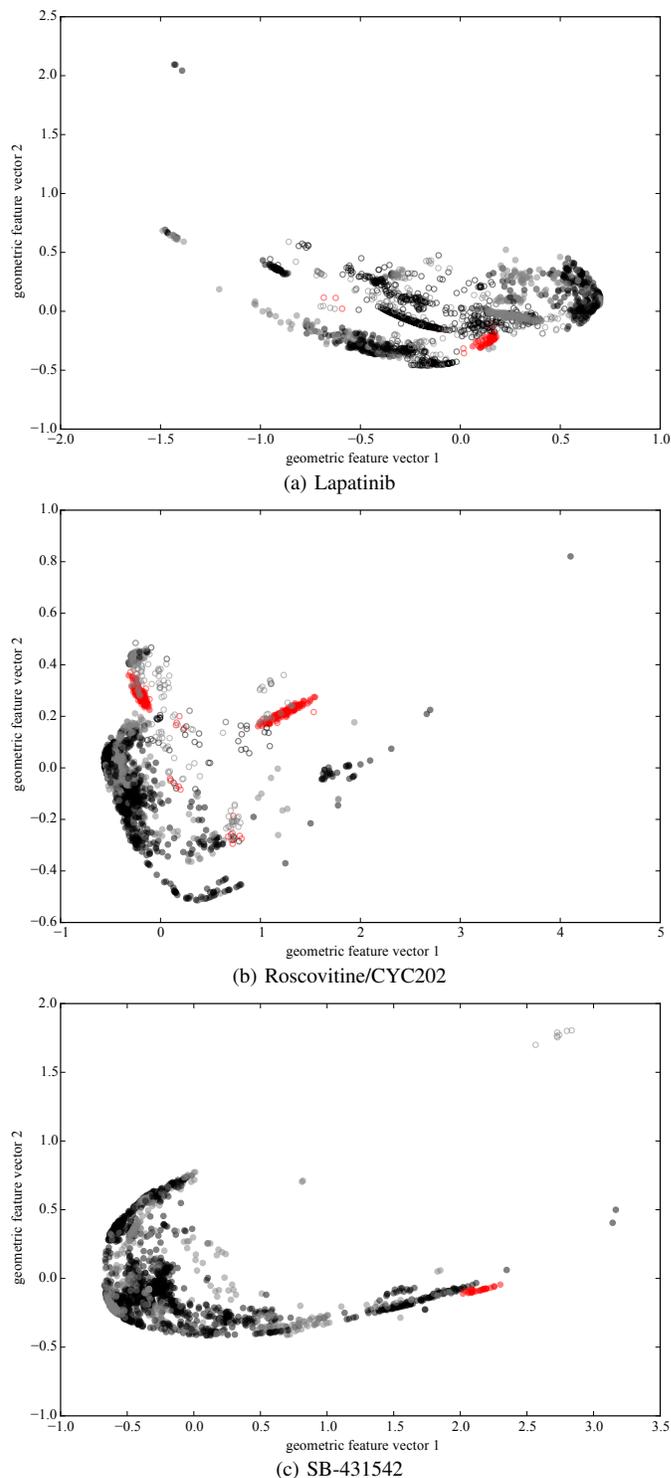
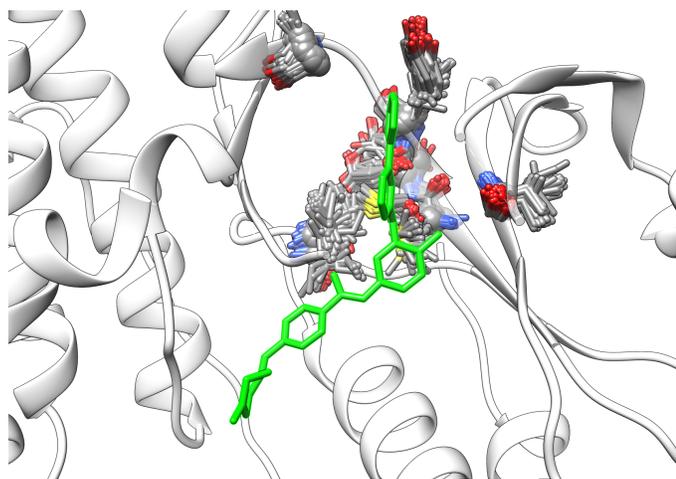
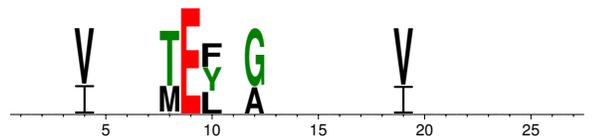


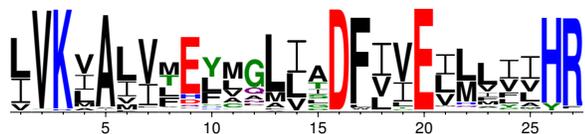
Fig. 2: Examples of the kind of clusterings selected by our algorithm. The axes correspond to the 2D, PCA-reduced feature vector representation of the pairwise distances between structures as described in Section III. Each point represents one structure. **Red**: known to bind, **black**: known to not bind, **gray**: binding affinity unknown. **Discs**: structures belonging to HPCs, **circles**: all other structures.



(a) P38 (grey ribbon) shown bound to Imatinib (green), PDB ID 3HEC. Superimposed are the SDPs, as determined by our algorithm, for all structures known to bind Imatinib, aligned with the corresponding residues of P38.



(b) Sequence logo for just the SDPs of structures known to bind Imatinib.



(c) Sequence logo for entire binding site using sequences from all 1,958 structures.

Fig. 3: Visual representations of a profile constructed by our algorithm: (a) the superposition of the selected residues for the structures that bind to Imatinib and (b) a sequence logo for those same structures.

sufficient for high selectivity. A high degree of structural conservation is also necessary, which appears to be the case here.

At a high level, residue positions that occur frequently in the profiles are often ones with known roles in inhibitor binding and selectivity determination. An example is the region which structural biologists term the “hinge” (position 9 in Fig. 3(c)), which binds to the adenine ring of the natural ATP substrate and is also used by the vast majority of kinase inhibitors. Another key residue frequently highlighted in logos is the “gatekeeper” residue [27] (position 8 in Fig. 3(c)); the size of this residue controls access to a secondary binding pocket and is a major determinant of selectivity. More specifically, the analysis for the kinase inhibitor Imatinib identifies these residues. In addition, several other residues in the profile are known to be associated with mutations that confer resistance to Imatinib.

### B. Coverage and Predictive Power of SDPs

Based on the set of SDPs we can (a) try to “recover” the labels of labeled structures that were not part of the selected

TABLE I: Coverage of labeled structures, number of predicted affinities for unlabeled structures, as well as sensitivity, specificity, precision, and accuracy for HPC-based prediction of binding affinity. Each row summarizes the average over all 38 ligands for the corresponding strategy.

Strategy	cov.	#pred.	sens.	spec.	prec.	acc.
1	83%	215	0.486	1.000	0.921	0.904
2	83%	215	1.000	0.887	0.783	0.929
3	15%	1,084	0.617	1.000	0.921	0.932
4	71%	364	1.000	0.900	0.806	0.937

TABLE II: Coverage of labeled structures, number of predicted affinities for unlabeled structures, as well as specificity, precision, and accuracy for HPC-based prediction of binding affinity as recovered from SDPs computed using our algorithm (with  $\lambda = 6$  and  $\delta = 16$ ). Sensitivity is equal to 1 in all cases.

Inhibitor	cov.	#pred.	spec.	prec.	acc.
ABT-869	86%	557	0.922	0.633	0.931
AMG-706	83%	558	0.928	0.707	0.938
AST-487	65%	426	0.661	0.806	0.859
AZD-1152HQA	85%	568	0.914	0.668	0.927
BIRB-796	67%	391	0.766	0.653	0.838
BMS-387032/SNS-032	96%	670	0.984	0.959	0.988
CHIR-258/TKI-258	81%	420	0.947	0.861	0.960
CHIR-265/RAF265	87%	473	0.960	0.801	0.966
CI-1033	77%	475	0.882	0.710	0.909
CP-690550	96%	629	0.989	0.736	0.989
CP-724714	99%	684	0.999	0.982	0.999
Dasatinib	83%	500	0.897	0.837	0.933
EKB-569	70%	474	0.876	0.688	0.902
Erlotinib	80%	532	0.902	0.693	0.920
Flavopiridol	80%	515	0.844	0.754	0.895
GW-2580	99%	677	1.000	1.000	1.000
GW-786034	79%	485	0.920	0.737	0.934
Gefitinib	81%	470	0.906	0.561	0.916
Imatinib	86%	587	0.936	0.590	0.941
JNJ-7706621	59%	356	0.580	0.704	0.790
LY-333531	83%	413	0.912	0.652	0.924
Lapatinib	99%	684	0.999	0.982	0.999
MLN-518	94%	659	0.989	0.808	0.989
MLN-8054	87%	493	0.948	0.766	0.956
PI-103	99%	654	0.999	0.988	0.999
PKC-412	54%	217	0.621	0.687	0.793
PTK-787	97%	664	0.999	0.974	0.999
Roscovitine/CYC202	98%	650	1.000	1.000	1.000
SB-202190	84%	500	0.929	0.815	0.946
SB-203580	69%	349	0.792	0.641	0.849
SB-431542	100%	670	1.000	1.000	1.000
SU-14813	71%	343	0.761	0.667	0.838
Sorafenib	70%	509	0.919	0.801	0.939
Staurosporine	91%	646	0.681	0.956	0.959
Sunitinib	61%	343	0.652	0.654	0.790
VX-680/MK-0457	78%	410	0.844	0.767	0.897
VX-745	85%	583	0.912	0.680	0.926
ZD-6474	87%	511	0.939	0.823	0.952
<b>average</b>	<b>83%</b>	<b>520</b>	<b>0.887</b>	<b>0.783</b>	<b>0.929</b>

HPCs and (b) predict labels for the unlabeled structures. There are at least four simple strategies to do this:

- 1) We could assume that the union of all true-HPCs contains all the structures that bind and that all others do not bind.
- 2) We could assume that the union of all false-HPCs contains all the structures that do not bind and all others

do bind.

- 3) We could omit the `false`-HPCs altogether from the input  $H$  to Algorithm 1 and select residue subsets based on large `true`-HPCs only. The labels are then recovered as in (1).
- 4) We could omit the `true`-HPCs altogether from the input  $H$  to Algorithm 1 and select residue subsets based on large `false`-HPCs only. The labels are then recovered as in (2).

Note that the SDPs computed with Algorithm 1 are the same in the first two strategies, but will generally look different when using strategies 3 and 4. We have evaluated each of these strategies on all 38 ligands. For each we can evaluate the coverage: the percent of known labels that are included in the HPCs. We can also count the number of *unlabeled* structures included in HPCs, which can be interpreted as the number of new binding affinities we can predict. For the first two strategies we get predictions for both binding and not-binding, while for the latter two we only get predictions for one type of affinity. Finally, we can calculate the usual statistical performance measures (sensitivity, specificity, precision, and accuracy) to measure how well the selected HPCs can predict binding affinity for all labeled structures. The results were computed with  $\lambda = 6$  and  $\delta = 16$  and are summarized in Table I. Note that that specificity is equal to 1 in strategies 1 and 3 by construction. Similarly, sensitivity is equal to 1 in strategies 2 and 4 by construction. In general, assuming that the union of all `true`-HPCs contains all the structures that bind (as is done in strategies 1 and 3) results in poor sensitivity. Strategy 2 seems to strike a good balance between sensitivity and specificity as well as between precision and accuracy. Strategy 4 performs even better than strategy 2, but provides poorer coverage.

The results in Table II show more detailed results for each ligand with strategy 2. While there is some variation among the inhibitors, the coverage is almost always very high. In cases where it is not, such as AST-487, JNJ-7706621 and Sunitinib, it is usually a inhibitor that binds to many different parts of the kinome tree (see kinome interaction maps in [25]). Finally, we analyzed the sensitivity to the parameter  $\delta$  and  $\lambda$ . As is shown in Tables III and IV, performance varies significantly with both  $\lambda$  and  $\delta$  (as is expected). However, even with very large values of  $\delta$ , the algorithm is still able to cover the vast majority of known binding affinities. Even more surprisingly, even when restricting SDPs to only  $\lambda = 3$  residues (corresponding to a *single* HPC), over 60% of the structures with known binding affinity are covered.

## VI. CONCLUSION

We have described a general method for identifying Specificity Determining Positions in families of related proteins. The method was shown to be very effective in identifying SDPs within the human kinome that help explain the binding affinity of 38 different inhibitors.

In ongoing work we are exploring the potential role of other residues identified by the structure-guided selection. Some

TABLE III: Sensitivity to the value of  $\lambda$  with  $\delta = 16$ . Each row represents an average over all 38 inhibitors.

$\lambda$	cov.	#pred.	spec.	prec.	acc.
3	62%	312	0.669	0.493	0.778
4	73%	419	0.781	0.661	0.864
5	79%	482	0.844	0.729	0.907
6	83%	520	0.887	0.783	0.929
7	86%	537	0.909	0.810	0.943
8	88%	554	0.921	0.838	0.951
9	89%	565	0.930	0.858	0.958

TABLE IV: Sensitivity to the value of  $\delta$  with  $\lambda = 6$ . Each row represents an average over all 38 inhibitors.

$\delta$	cov.	#pred.	spec.	prec.	acc.
1	85%	587	0.904	0.820	0.941
2	85%	580	0.903	0.817	0.940
4	85%	565	0.900	0.812	0.938
8	84%	547	0.895	0.800	0.935
16	83%	520	0.887	0.783	0.929
32	81%	490	0.871	0.723	0.916
64	78%	456	0.848	0.658	0.898
128	74%	413	0.817	0.612	0.876

of these are not in direct contact with the inhibitor but may be involved indirectly through, for example, influencing the conformation or flexibility of the protein. This would be a significant benefit, as such residues are difficult to identify by other means. Not only could this potentially provide a new insight into the structural biology of kinases, but such knowledge may be helpful in the design of inhibitors with novel, or improved, selectivity profiles.

In prior work [28] we have demonstrated that the addition of homology models leads to an improvement in the prediction of binding affinity. Homology models can fill in gaps in structural coverage, thereby potentially eliminating “accidental” HPCs and create new ones. In future work we plan to investigate whether homology models can provide similar benefits in the identifications of SDPs.

## ACKNOWLEDGMENT

The authors wish to thank Drew Bryant. Without his work on creating the CCORPS software infrastructure and preparing the human kinase dataset for processing with CCORPS the results presented here would not have been possible.

## REFERENCES

- [1] D. H. Bryant, M. Moll, P. W. Finn, and L. E. Kavasaki, “Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome,” *PLOS Computational Biology*, vol. 9, no. 6, p. e1003087, Jun. 2013.
- [2] Y. Liu and N. S. Gray, “Rational design of inhibitors that bind to inactive kinase conformations,” *Nat Chem Biol*, vol. 2, no. 7, pp. 358–64, Jul 2006.
- [3] D. Kuhn, N. Weskamp, E. Hüllermeier, and G. Klebe, “Functional classification of protein kinase binding sites using Cavbase,” *ChemMedChem*, vol. 2, no. 10, pp. 1432–47, Oct. 2007.

- [4] J. A. Bikker, N. Brooijmans, A. Wissner, and T. S. Mansour, "Kinase domain mutations in cancer: implications for small molecule drug design strategies," *J Med Chem*, vol. 52, no. 6, pp. 1493–509, Mar 2009.
- [5] F. Milletti and J. C. Hermann, "Targeted kinase selectivity from kinase profiling data," *ACS Med Chem Lett*, vol. 3, no. 5, pp. 383–386, May 2012.
- [6] O. A. Gani, B. Thakkar, D. Narayanan, K. A. Alam, P. Kyomuhendo, U. Rothweiler, V. Tello-Franco, and R. A. Engh, "Assessing protein kinase target similarity: Comparing sequence, structure, and cheminformatics approaches," *Biochim Biophys Acta*, May 2015.
- [7] P. W. Finn and L. E. Kavraki, "Computational approaches to drug design," *Algorithmica*, vol. 25, no. 2, pp. 347–371, Jun. 1999.
- [8] O. V. Kalinina, A. A. Mironov, M. S. Gelfand, and A. B. Rakhmaninova, "Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families," *Protein Sci*, vol. 13, no. 2, pp. 443–56, Feb 2004.
- [9] S. Chakrabarti and C. Lanczycki, "Analysis and prediction of functionally important sites in proteins," *Protein Science*, vol. 16, no. 1, p. 4, 2007.
- [10] J. A. Capra and M. Singh, "Characterization and prediction of residues determining protein functional specificity," *Bioinformatics*, vol. 24, no. 13, pp. 1473–1480, Jul. 2008.
- [11] F. Pazos, A. Rausell, and A. Valencia, "Phylogeny-independent detection of functional residues," *Bioinformatics*, vol. 22, no. 12, pp. 1440–1448, Jun. 2006.
- [12] A. Rausell, D. Juan, F. Pazos, and A. Valencia, "Protein interactions and ligand binding: from protein subfamilies to functional specificity," *Proc Natl Acad Sci U S A*, vol. 107, no. 5, pp. 1995–2000, Feb 2010.
- [13] J. Mok, P. M. Kim, H. Y. K. Lam, S. Piccirillo, X. Zhou, G. R. Jeschke, D. L. Sheridan, S. A. Parker, V. Desai, M. Jwa, E. Cameroni, H. Niu, M. Good, A. Remenyi, J.-L. N. Ma, Y.-J. Sheu, H. E. Sassi, R. Sopko, C. S. M. Chan, C. De Virgilio, N. M. Hollingsworth, W. A. Lim, D. F. Stern, B. Stillman, B. J. Andrews, M. B. Gerstein, M. Snyder, and B. E. Turk, "Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs," *Sci Signal*, vol. 3, no. 109, p. ra12, 2010.
- [14] I. Halperin, D. S. Glazer, S. Wu, and R. B. Altman, "The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications," *BMC Genomics*, vol. 9 Suppl 2, p. S2, 2008.
- [15] T. Liu and R. B. Altman, "Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding," *PLoS Comput Biol*, vol. 7, no. 12, p. e1002326, Dec 2011.
- [16] B. Y. Chen and B. Honig, "VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity," *PLoS Comput Biol*, vol. 6, no. 8, p. e1000881, 2010.
- [17] S. L. Kinnings and R. M. Jackson, "Binding site similarity analysis for the functional classification of the protein kinase family," *Journal of Chemical Information and Modeling*, vol. 49, no. 2, pp. 318–329, 2009.
- [18] R. C. de Melo-Minardi, K. Bastard, and F. Artiguenave, "Identification of subfamily-specific sites based on active sites modeling and clustering," *Bioinformatics*, vol. 26, no. 24, pp. 3075–3082, Dec. 2010.
- [19] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D281–8, 2008.
- [20] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLoS Comput Biol*, vol. 4, no. 1, p. e10, Jan 2008.
- [21] C. Schalon, J.-S. Surgand, E. Kellenberger, and D. Rognan, "A simple and fuzzy method to align and compare druggable ligand-binding sites," *Proteins*, vol. 71, no. 4, pp. 1755–1778, Jun 2008.
- [22] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proc Natl Acad Sci U S A*, vol. 105, no. 14, pp. 5441–5446, Apr. 2008.
- [23] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash algorithm for substructure matching," *BMC Bioinformatics*, vol. 11, no. 555, Nov. 2010.
- [24] I. T. Jolliffe, *Principal Components Analysis*. New York: Springer-Verlag, 1986.
- [25] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar, "A quantitative analysis of kinase inhibitor selectivity," *Nat Biotechnol*, vol. 26, no. 1, pp. 127–132, Jan 2008.
- [26] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, no. 6, pp. 1188–1190, Jun 2004.
- [27] D. Huang, T. Zhou, K. Lafleur, C. Nevado, and A. Caflisch, "Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis," *Bioinformatics*, vol. 26, no. 2, pp. 198–204, 2010.
- [28] J. Chyan, M. Moll, and L. E. Kavraki, "Improving the prediction of kinase binding affinity using homology models," in *Computational Structural Bioinformatics Workshop at the ACM Conf. on Bioinf., Comp. Bio. and Biomedical Informatics*, Washington, DC, Sep. 2013, pp. 741–748.