Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Molecular Graphics and Modelling 24 (2006) 278–295
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Kinase inhibitor recognition by use of a multivariable QSAR model

D.G. Sprous *, John Zhang, Lei Zhang, Zhaolin Wang, M.A. Tepper

CytRx Laboratories, 1 Innovation Drive, Worcester MA 01605, USA

Received 27 June 2005; received in revised form 12 September 2005; accepted 13 September 2005
Available online 25 October 2005

Abstract

We have applied a retrosynthetic program to determine the scaffold and R-group chemical space seen within a library of known kinase inhibitors and nonkinase druglike molecules. Comparison of the differences quickly revealed that kinase inhibitors are distinct in several chemical fragment and physical properties. We then applied these descriptors in a multivariable quantitative structure–activity relationship (QSAR) model with the goal to distinguish kinase inhibitors from nonkinase druglike molecules. This model is heuristic in that it was trained over a dataset of 258 known kinase inhibitors and 230 nonkinase drug molecules. The final model recognized 98% of the training set as being kinase inhibitors and had a false positive rate of 15%. This trait for false positives was accepted out of a desire to maintain diversity and not miss possible good kinase inhibitors for screening. The model was validated by reserving a portion of the datasets as test sets, which were not included in the QSAR model building stage. This was done repetitively for different percentiles of the total dataset population. It was seen that model recognition and false positive were only slightly damaged well down to a 70% reserve (30% dataset used for QSAR model training while 70% used for reserve test set). Beyond 70%, the QSAR models were inconsistent, signifying that the training sets were inadequately diverse to represent the greater reserve test sets. We applied this model to evaluate the commercial kinase libraries available from Asinex, BioFocus, ChemDiv and LifeChemicals to facilitate purchase decisions for compounds for HTS for lead compounds. We observed that there are significant differences in populations of recognizable kinase inhibitors across the vendors analyzed, with BioFocus showing the greatest population of kinase like molecules.

Keywords: QSAR model; Kinase inhibitor; Retrosynthetic program

1. Introduction

In principle, high-throughput screening of diverse compound libraries can provide lead compounds for further optimization. Robotic systems permit the assay and analysis of literally hundred thousands to millions of compounds in a matter of days. In practice, protein quantities may be precious or assays may not be amenable for use in a high-throughput regime. This situation leads to a motive for smaller libraries, which are biased to the particular target of interest. For some target classes, such as GPCR's and kinases, there is a common similarity of active sites. This similarity of the active sites dictates a similarity of inhibitors, which can be quickly perceived in even a short inspection of the literature. With the motive available and the similarity readily seen, it is natural that focused libraries have been developed for both these classes. The present paper concerns the development, testing and use of a QSAR model to recognize compounds that are more likely to inhibit kinases than ordinary druglike compounds.

Roughly 500 protein kinases [1,2] exist. These proteins are, by definition, proteins that phosphorylate other proteins. The vast majority of these are recognizable at a sequence level by a conserved eukaryote protein kinase catalytic domain [1,2]. Members of this protein class consume ATP to add phosphates to another protein, many which are themselves kinases, toggling the target protein between forms that differ in activity. All protein kinases share common sequence and structural features within the ''Hinge'' region where ATP binds [3–5]. Specific kinase dependent pathways that are therapeutically relevant include differentiation, apoptosis and cell cycle control. Malfunctions in the kinase dependent cellular systems leads to a variety of problems including nonexclusively cancer diabetes, and inflammation. Protein kinases' phenomenological place as activity modulators for entire biological cascades make them attractive targets for pharmaceutical research [5,6]. To date, only four compounds have reached commercial use (Fasudil for rho-dependent kinase, Rapamycin for TOR, Gleevec for BCR-Abl, and Iressa for EGFR [5,7–14]), though

* Corresponding author. Tel.: +1 5087673861; fax: +1 5087673862.
E-mail address: dsprous@cytrx.com (D.G. Sprous).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.09.004

many more are in clinical development and it is impossible to determine how many are in preclinical research [5].

However, kinase sequence and structural similarity is a cause for concern [3,5,15,16]. Indeed, in the early 1990's, it was widely assumed that kinases were not viable pharmaceutical targets due to their inherent similarity. Nucleotide analogs, directly competing with the adenine and ribose rings of ATP, proved excellent kinase inhibitors for the entire kinase class but were not viable as therapeutics since they had dramatic effects over too numerous cellular processes. In short, nucleotide analogs were chainsaws where scalpels were needed. However, nonnucleotide analogs were developed that are pharmaceutically viable. These inhibitors do in fact, frequently compete with the adenine ring in the so called hinge region. However, these compounds also reach into an irregular shaped, largely hydrophobic pocket. Many authors have partitioned this pocket into different zones. However, the border between ATP and this pocket is defined by a gatekeeping residue (typically methionine, serine or phenylalanine), frequently a lysine and Mg^{2+} [3,4]. The exact dimensions and residue content of this zone differ between specific kinases. By reaching into this pocket, specificity is achievable, though there is still a similarity of the overall active site topology, which in turns dictates similarity in the chemical structure of small molecules able to fit the active site. This similarity of compounds has led to the development of kinase focus libraries [6]. These libraries have been developed both for internal research and external sale.

Kinase focus libraries are developed based on combinatorial chemistry of privilege scaffolds and sidechains. Identification of these special scaffolds and sidechains has been done by (or by combination of) literature inspection, virtual docking against consensus sets of kinase proteins, pharmacophore based screening and similarity searching. Despite the apparent diversity of techniques listed above, the majority of strategies have common goals which include recognizing (1) chemical fragments that will be able to recognize the hydrogen bonding atoms on the protein that are paired with the adenosine ring, (2) other chemical fragments which will fit in the more variable hydrophobic pocket and (3) recognizing linker chemical fragments which will be able to pass the gatekeeper residue and connect the two. The linker fragments must themselves show certain diversity since the gatekeeper residue will vary and the exact required vectors to connect the occupant of the adenosine and hydrophobic pocket will vary dependent on exact nature of the kinase and the chemical fragments in each site. Though no work has been done to our knowledge concerning developing ATP noncompetitive inhibitor kinase focus libraries, this would simply modify the goals above to focus exclusively on the variable hydrophobic pocket.

Numerous papers have documented application of similar strategies to design libraries targeted to a single kinase or a small number of closely related kinases (see the reviews in [17] and references therein). However, though creation of these libraries is a common practice for both compound vendors and pharmaceutical chemistry division, details and publications are relatively sparse for these endeavors with some key exceptions [18,19]. For compound vendors, this may be motivated by a fear that any provided metric could be used in a negative way by an evaluating customer. Further, there is a lack of freely deposited databases, which would enable interested parties to develop the ability to estimate metrics for kinase libraries. The present paper was stimulated by the desire to develop a strategy to estimate the kinase inhibitor quantity available in marketed vendor libraries. To satisfy this desire, two questions had to be addressed. The first question was ''What is the scaffold and sidechain group content for kinase inhibitors and how does this differ from other druglike molecules?''. After this question was evaluated, the second question was ''What type of physical–chemical descriptors can be employed to recognize kinase inhibitors from other druglike compounds?''.

To answer these questions, we first assembled from literature [20] a dataset of potent known kinase inhibitors and know drug molecules [21,22]. Thereafter, we applied a retrosynthetic program [23] to convert our assembled databases into scaffold and sidechain fragment libraries. By application of simple subset operations, were we able to create subsets that clarified what was common versus unique in kinase inhibitors versus regular druglike molecules for both the scaffolds and the sidechains. The chemical fragment content itself is applicable as descriptors to develop models to recognize kinase inhibitors from nonkinase druglike molecules. However, we recognized quickly that such a model would have a limited use radius since its chemical space was so exactly defined by our too finite training datasets. Reserve test sets were poorly predicted with these models, signifying a narrow use of the models outside of the training set. To address this problem, we posed the question of what is common to the unique fragments in both the kinase and nonkinase sets. This exercise produced a list of numerous ''traits'' for each set. These traits were developed into general definitions for descriptors, which proved to differ by degrees for kinase specific versus other pharmaceutical compounds. The performance of these descriptors in our a kinase inhibitor recognition QSAR model proved to be excellent, with strong abilities to predict correctly compounds in reserve test sets as being either kinase specific inhibitors or nonkinase specific inhibitors.

2. Materials and methods

This section details datasets employed and programs used. All datasets (excluding vendor datasets), SYBYL programming language (SPL) scripts and final QSAR models are provided as supporting information. All calculations were performed on a LINUX based workstation within SYBYL [24].

2.1. Datasets

2.1.1. Training datasets

This modeling relied on training sets of known kinase inhibitors and known small molecule pharmaceutical compounds (provided as SD-files as supporting information). The 258 kinase inhibitors were derived from a series of review papers [20]. The list of 230 known pharmaceutical compounds was assembled from the available dataset of Yoshida & Topliss [21], which is nearly identical to the dataset of Oprea &

Table 1
Commercial datasets investigated and populations

Vendor	Diversity library population (K)	Kinase library population
Asinex	190	4307
BioFocus		16,000
ChemDiv	467	9879
ChemBridge	427	
LifeChemicals	150	40,256

Gottfries [22]. The inhibitors are not focused on any one specific kinase group. Further, the assays used to determined potency were naturally not uniform. Likewise, the dataset of nonkinase inhibitors has no specific target or therapeutic area. For this reason, the compounds are entered with a single dependent variable "IS_KINASE_INHIBITOR", which has a value of 0 for a kinase inhibitor and a value of 1 for a nonkinase inhibitor.

2.1.2. Vendor datasets analyzed

Datasets for commercially offered compounds were obtained from Asinex [25], BioFocus [26], ChemBridge [27], ChemDiv [28] and LifeChemicals [29] (Table 1). These datasets are typically available on request from the vendors.

2.2. CLRP: CytRx laboratories retrosynthetic program

This program is an SPL script originally developed by Dr. Jon Swanson at Tripos Inc. and extended at CytRx Laboratories. The script is 1300 lines. This retrosynthetic program is based on the work of Lewell et al. [23]. Like the original Lewell RECAP program, CLRP breaks compounds according to synthetically specified cleavage rules (see Fig. 1). These cleavage rules are implemented as SLN's. In addition to this original fragmentation function, CLRP distinguishes between scaffolds and sidechains. Scaffolds are defined as fragments with more than a single point where a cut was made in producing the fragment. Sidechains specifically have only one point where a cut was made. The definition of cleavage patterns is tied directly to commonly known synthetic chemistry patterns. The scaffold and sidechain separation is more arbitrary and was developed to organize fragments in a manner that would match medicinal chemical intuition. The scaffold definition allows a compound to have more than a

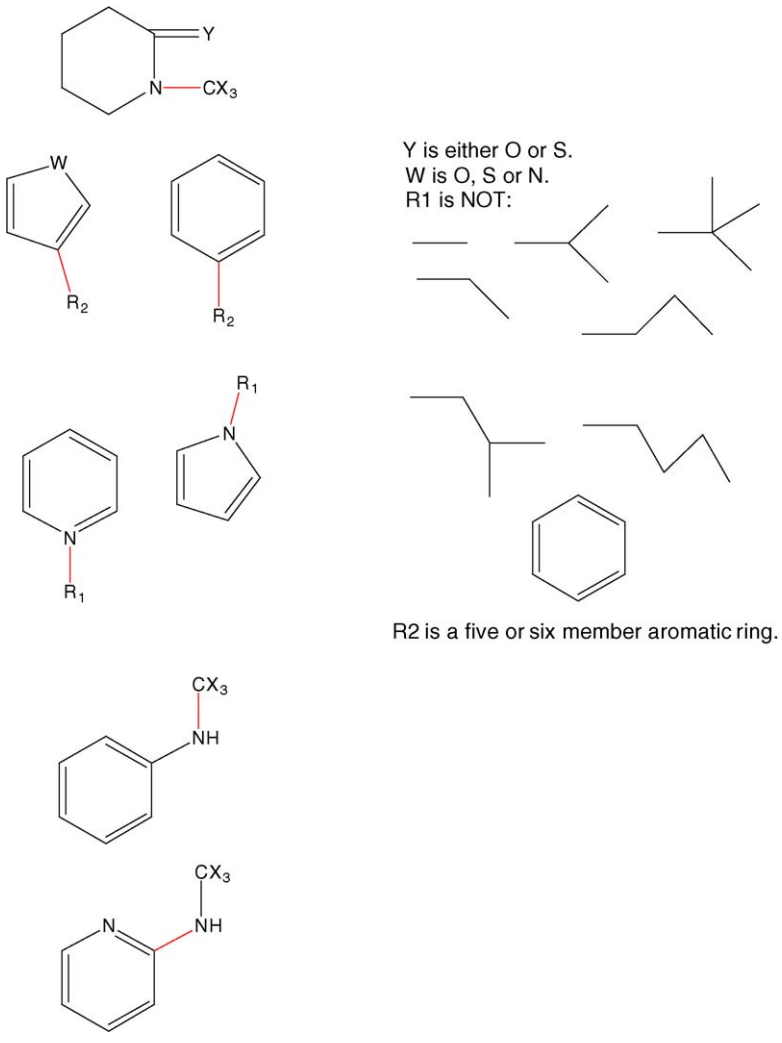


Fig. 1. Rules for virtual cleavage of compounds. Red bonds and atoms denote bonds and atoms lost in application of the cleavage rules.

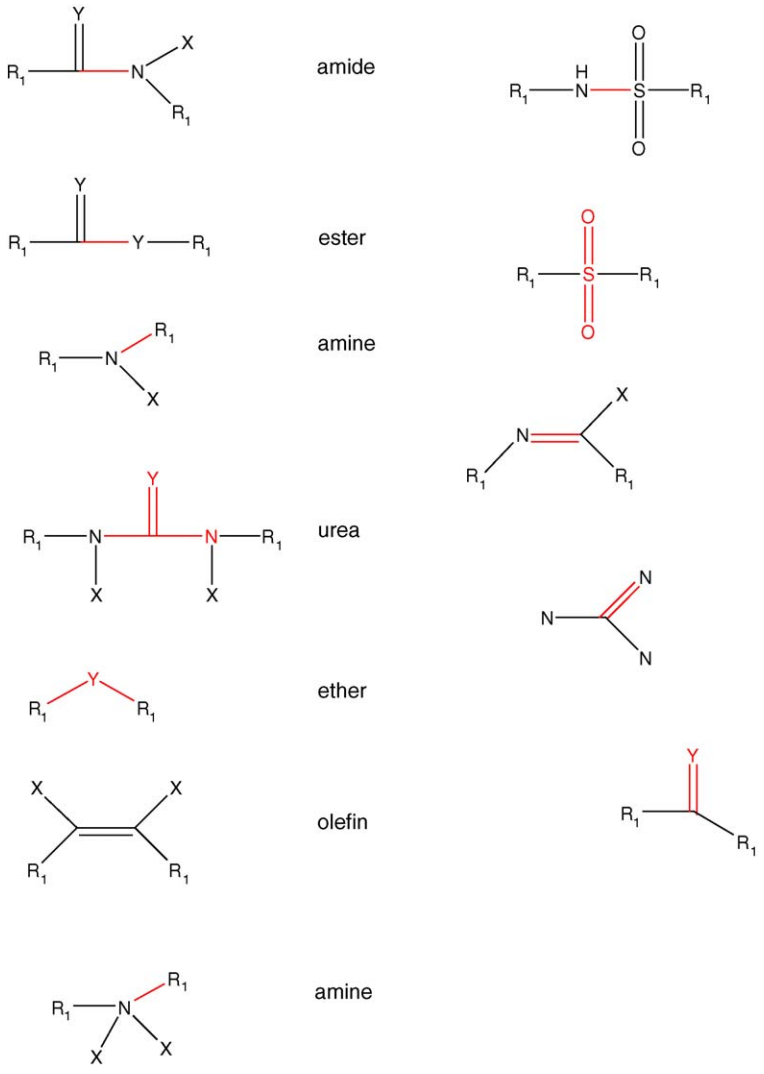


Fig. 1. (*Continued*).

single scaffold, which may seem odd to most chemists' intuition. Still, we consider the demarcation between sidechains and scaffolds to be useful, albeit arbitrary.

2.3. Descriptors employed

The definition for the descriptors employed in final models is presented in Table 2. The majority of the descriptors are chemical fragment in origin and the definition employed is presented as SLN notation. Some of the descriptors are native SYBYL descriptors and are denoted by "SYBYL" in the table. The descriptor "Flex" is a ratio of the rotable bond count to the total bond count. We had alternatively considered "Flex" to be the ratio of the rotable bond count to the molecular weight. However, this leads to comparison problems since a change of O to S or Cl to Br has an undesired dramatic impact on a metric, which should be the relative flexibility of the molecule. These descriptors are similar in spirit to those commonly seen in the oral bioavailability or "druglike" scientific journal article genre [20,21,30–33]. SPL scripts for generating these are provided as supplemental information from the author.

2.4. QSAR model generation, validation and application

2.4.1. QSAR model generation

Partial least squares [34] were used to develop our QSAR models. The proper number of components was taken as being when the standard error was at the first minima. Due to the simple Boolean nature of our data (compounds are or are not kinase inhibitors and assigned a value of 0 or 1 to reflect this), high values for the Pearson's correlation coefficient are not the primary goal. The real success criterion is the QSAR models performance in recognition versus false positives. Recognition is defined as assigning a score between −1.00 and 0.50 to a compound known to be a kinase and a false positive is defined as assigning a score of −1.00 to 0.50 to a nonkinase compound. This restriction to a finite range also limits the models application to an area of space, which is more likely to be within the chemical space over which the QSAR model was trained.

2.4.2. Validation

Our datasets are of adequate size that most compounds have one or more close "sibling(s)" compounds, which are quite

Table 2
Descriptors devised based on consideration of the chemical content of the two datasets

TAG	SLN or formula
Molecular weight	SYBYL
Atom count	SYBYL
Bond count	SYBYL
Hydrophobe	SYBYL
Ring count	SYBYL
TPSA	SYBYL
Rotable bonds	SYBYL
Donor	SYBYL
Acceptor	SYBYL
Flex	Rotbond/Bondcount
IGNR	C:N:C
Aromatic carbons	Any:C:Any
Aromatic nitrogens	Any:N:Any
Aromatic oxygens	Any:O:Any
Aromatic sulfurs	Any:S:Any
Aryl-amines	Any:C(–N(Any)(Any)):Any
Halogen	Any[is = F,Cl,Br,I]
Saturated halogen	AnyC(Any)(Any)Any[is = F,Cl,Br,I]
Aromatic halogen	Any:C(Any[is = F,Cl,Br,I]):Any
F	F
Saturated F	AnyC(Any)(Any)F
Aromatic F	Any:C(F):Any
CL	Cl
Saturated CL	AnyC(Any)(Any)Cl
Aromatic CL	Any:C(Cl):Any
BR	Br
Saturated Br	AnyC(Any)(Any)Br
Aromatic Br	Any:C(Br):Any
I	I
Saturated I	AnyC(Any)(Any)I
Aromatic I	Any:C(I):Any
Five member aromatic rings	Any[1]:Any:Any:Any:Any:@1
Six member aromatic rings	Any[1]:Any:Any:Any:Any:Any:@1
Branched aromatic carbon	Any:C(Any[is = C,N,O,S]):Any
Branched saturated carbon	Any[is = C,N,O,S]–C(Any[is = C,N,O,S])Any[is = C,N,O,S]
Branched aromatic nitrogen	Any:N(Any[is = C,N,O,S]):Any
Branched saturated nitrogen	Any[is = C,N,O,S]–N(Any[is = C,N,O,S])–Any[is = C,N,O,S]

similar in overall structure and properties. This situation makes application of Leave-One-Out-crossvalidation inapplicable [35]. To address the need for a validation process, we systematically removed compounds from the training set and placed them in reserve sets, which were then scored according to the QSAR model. To develop these sets in an unbiased manner, random numbers between 0.00 and 1.00 were generated for each compound according to three different seeds. Thereafter, all compounds below some specified value were used as members of the training set to generate a QSAR model. The model was then applied to predict the test set and the values for recognition and false positives where recorded. This was repeated for each possible random number seed event, to verify the range of deviation seen for the models when this exclusion process was applied. We were interested most critically in the worst case scenario performance for the model rather than the average, since we felt that reflected the most conservative estimate for the model's use radius.

2.4.3. Application

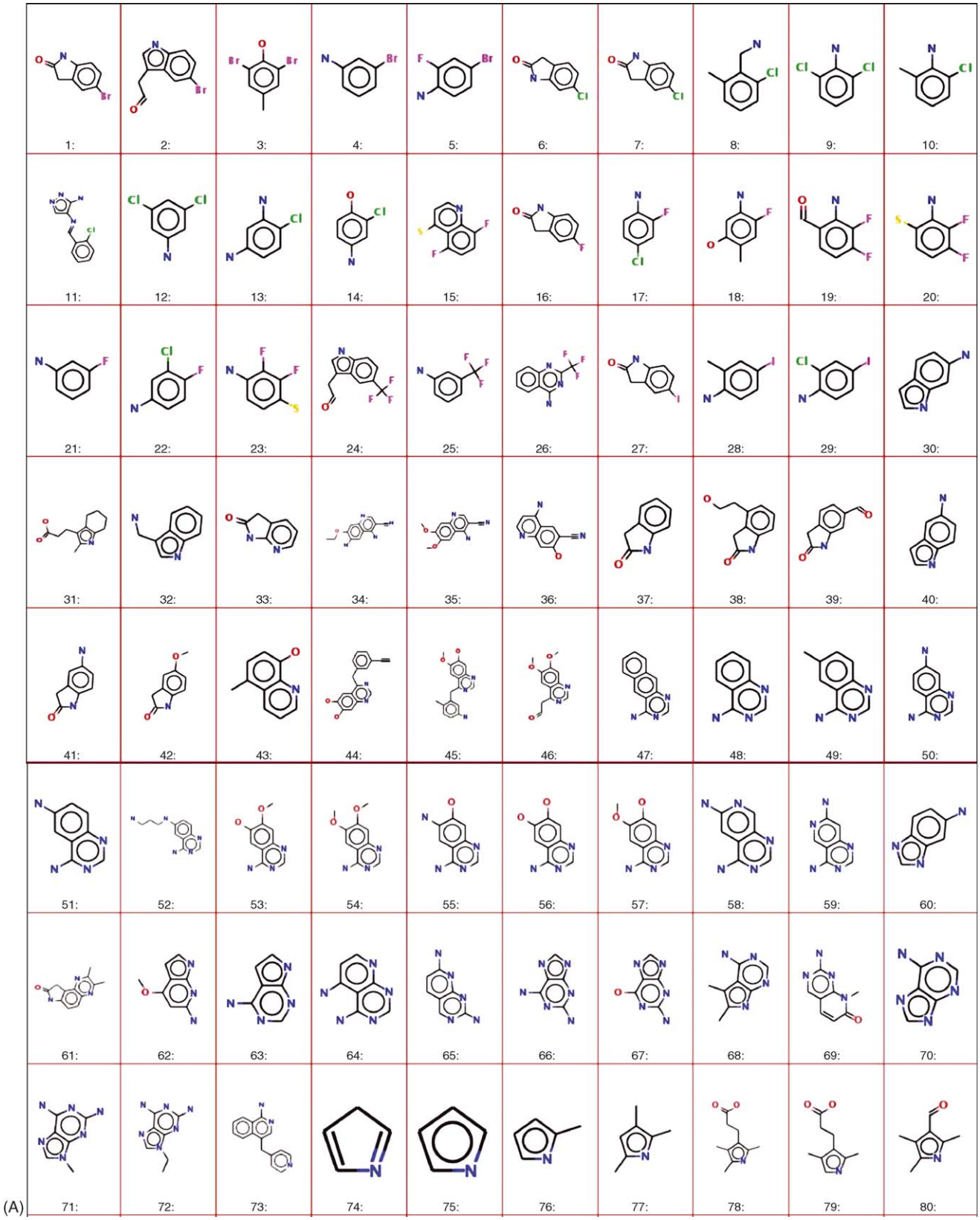
The QSAR model developed over the full available set of kinase inhibitors and nonkinase inhibitors was used for evaluation of commercially purchasable compounds. Each analysis consisted of using the QSAR model to assign a score to each compound. Thereafter, all compounds that were assigned a value between −1.0 and 0.5 were deemed to be both within the model use space and to be kinase inhibitors. There are compounds that are scored more negatively than this, but these must be considered outside of the QSAR models use radius and are not included as kinase inhibitors.

3. Results and discussion

Our first attempt to determine what is unique to kinase inhibitors started by application of the CLPR program to the kinase inhibitor dataset and to the Yoshida–Topliss dataset. We performed set selection to partition the fragments into three primary classes: (1) those unique to kinase inhibitors, (2) those unique to the Yoshida–Topliss drug compounds and (3) those common to both kinase inhibitors and to the Yoshida–Topliss drugs compounds (see Figs. 2–4). This was done partly to develop a list of privileged fragments. The set of commons turned out to be relatively small, which was taken as a signal that it would be possible to develop a recognition function for discriminating between kinase inhibitors and nonkinase drugs.

The kinase inhibitors are seen to have 185 unique scaffolds and 155 unique sidechains. The Yoshida–Topliss dataset has 132 unique scaffolds and 157 unique sidechains. For the kinase inhibitors, all but 10 scaffolds bear aromatic rings while for the Yoshida–Topliss dataset 36 lack aromatic rings. Thirty-five kinase scaffolds are fused ring structures in the kinase dataset. The number of fused rings is comparable for Yoshida–Topliss with 22 fused rings. The amount of aromatic systems seems roughly equal from these simple counts but the Yoshida–Topliss fragments are clearly richer in mixes of saturated bonds and aromatic systems while the aromatic content in the kinase fragments is nearly to the exclusion of any saturated system. A key visual trait seen in the figures is the large number of complex aromatic substitutions in kinase inhibitors. There is an abundance of multiple nitrogen substitutions and multiple heteroatom substitutions. The fragments in Fig. 3 are less likely to be halogenated. The two datasets have in common 21 scaffolds and 34 sidechains. These fragments are of a relatively simple nature.

We did attempt to develop QSAR models, which used the occurrences of the unique scaffolds and sidechains as descriptors. These models performed well over the training set with near perfect recognition rates and no false positives. However, the models were overfitted and had little ability to predict outside of their training sets. A preliminary test was performed where the scaffolds and sidechains were ''simplified'' by adding instructions to cut methylene-aromatic bonds



(A)

Fig. 2. (A) Scaffolds and (B) sidechains found in the kinase inhibitor database but not found in the Yoshida–Topliss dataset.

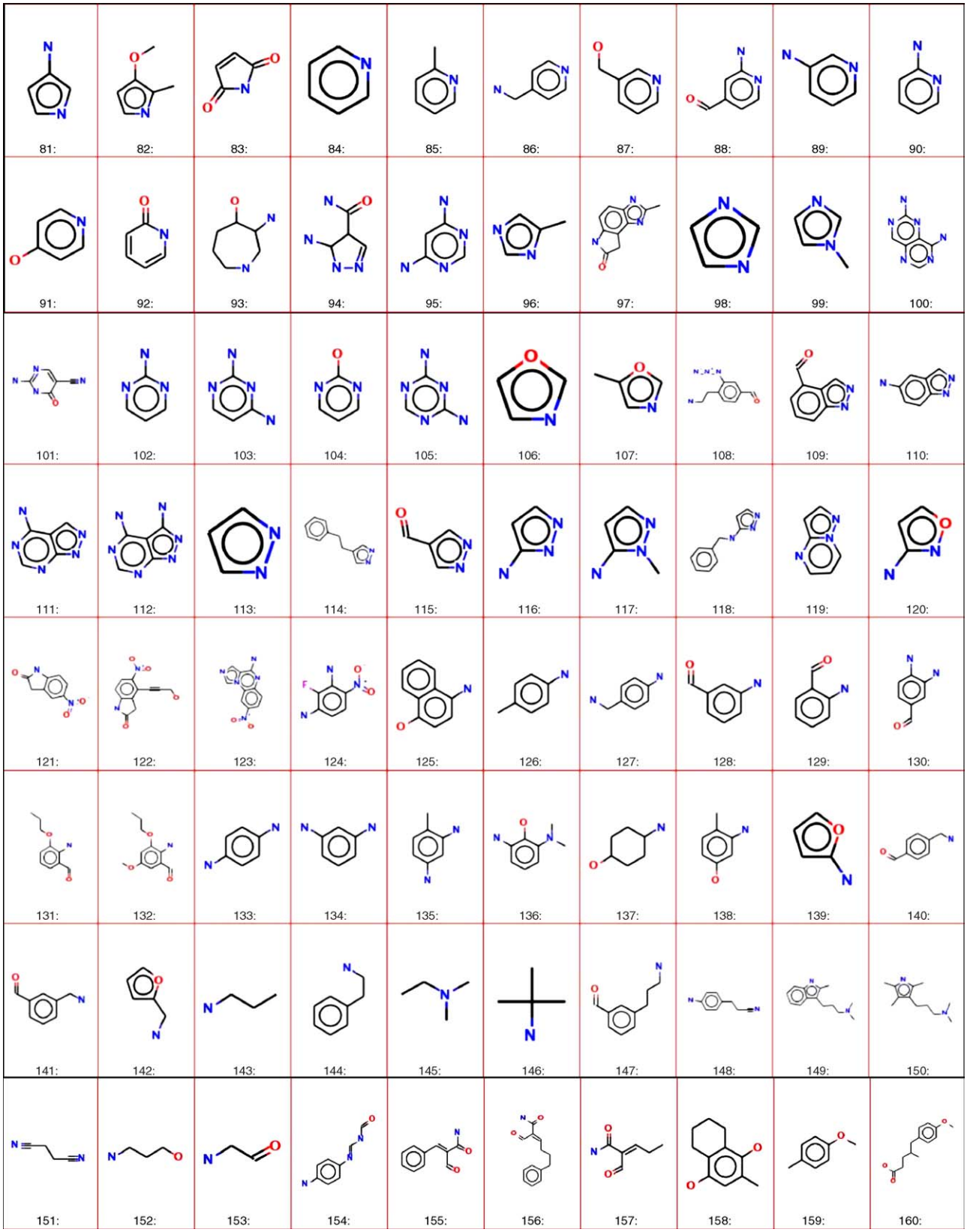


Fig. 2. (*Continued*).

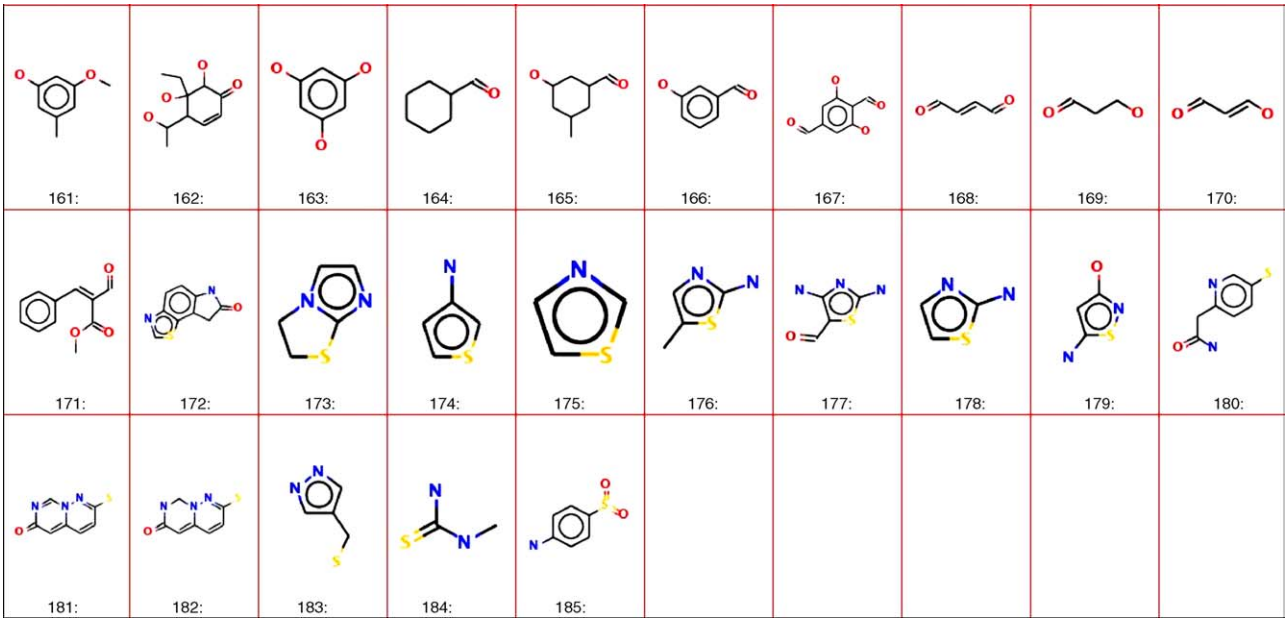


Fig. 2. (*Continued*).

and to eliminate halogens. This produced a smaller number of scaffolds and sidechains which when applied as QSAR descriptors performed better at prediction outside of test sets.

However, the use radius still seemed too narrow so analysis was undertaken to develop insights to what is unique in a physical and chemical space about scaffolds or sidechains in both groups. These insights eventually led to proposals for descriptors that allowed QSAR models to operate comfortably outside of the training sets. Numerous descriptors applied were know from the ''druglike compound'' scientific journal genre and include the first nine entries seen in Table 2. These are simple metrics are atomic mass, TPSA [30], counts for atoms, bonds, rotatable bonds, donors, acceptors, hydrophobes and rings. The more crucial descriptors though included those devised from the qualitative description of the differences between kinase fragments and Yoshida–Topliss fragments. These are outlined in Table 2 starting with the 10th entry and are essentially exercises in specialized atom typing and counts.

As that exclusion of the three terms ''Atom Count'', ''Bond Count'' and ''Molecular Weight'' was seen to have no detrimental effect, our final models are based on a list of 34 descriptors. The QSAR model built on these was constructed by the use of partial least squares with four components, which was seen to be the first minima in the standard error. The four component model has the following statistics: $R^2 = 0.63$, F-value = 237, S.E. = 0.30. Exclusion of the three terms ''Atom Count'', ''Bond Count'' and ''Molecular Weight'' was seen to have no detrimental effect and these were eliminated from final models. All other terms, when eliminated, resulted in a detrimental effect to the QSAR models.

The models performance in terms of ''Recognition'' and ''False Positive'' rates is detailed in Table 3. The QSAR model itself when developed over the whole training set and applied to only the training set has a 98% recognition and a 15% false positive rate. To determine the protocols ability to function outside of the training set, the data was partitioned into ''Training Set'' and ''Reserve Test Set'' systematically. As was detailed in Section 2, a different random partition was selected three times at a specified population ratio. Thereafter, the model was developed over the ''Training Set'' and applied to predict the ''Test Set''. Averages for the three trials and ''WCS – Worse Case Scenarios'' were collected. Initial trials with high population ratios favoring the Training Set were very encouraging, which stimulated a jump to lower ratios quickly.

In Table 3, it is seen that for the case set where the ''Reserve'' is 70% of the total (R1_70, R2_70 and R3_70),

Table 3
Performance of the QSAR model across various trials

Trial ID	Size training set (%)	Size reserve set (%)	Performance rates (%)	
			Recognition	False positive
	100	0	98	15
R1_70			88	11
R2_70	30	70	86	7
R3_70			90	10
	AVG		88	9
	WCS		86	11
R1_80			87	10
R2_80	20	80	75	6
R3_80			85	9
	AVG		82	8
	WCS		75	10
R1_90			86	11
R2_90	10	90	69	11
R3_90			84	11
	AVG		80	11
	WCS		69	11

(B)

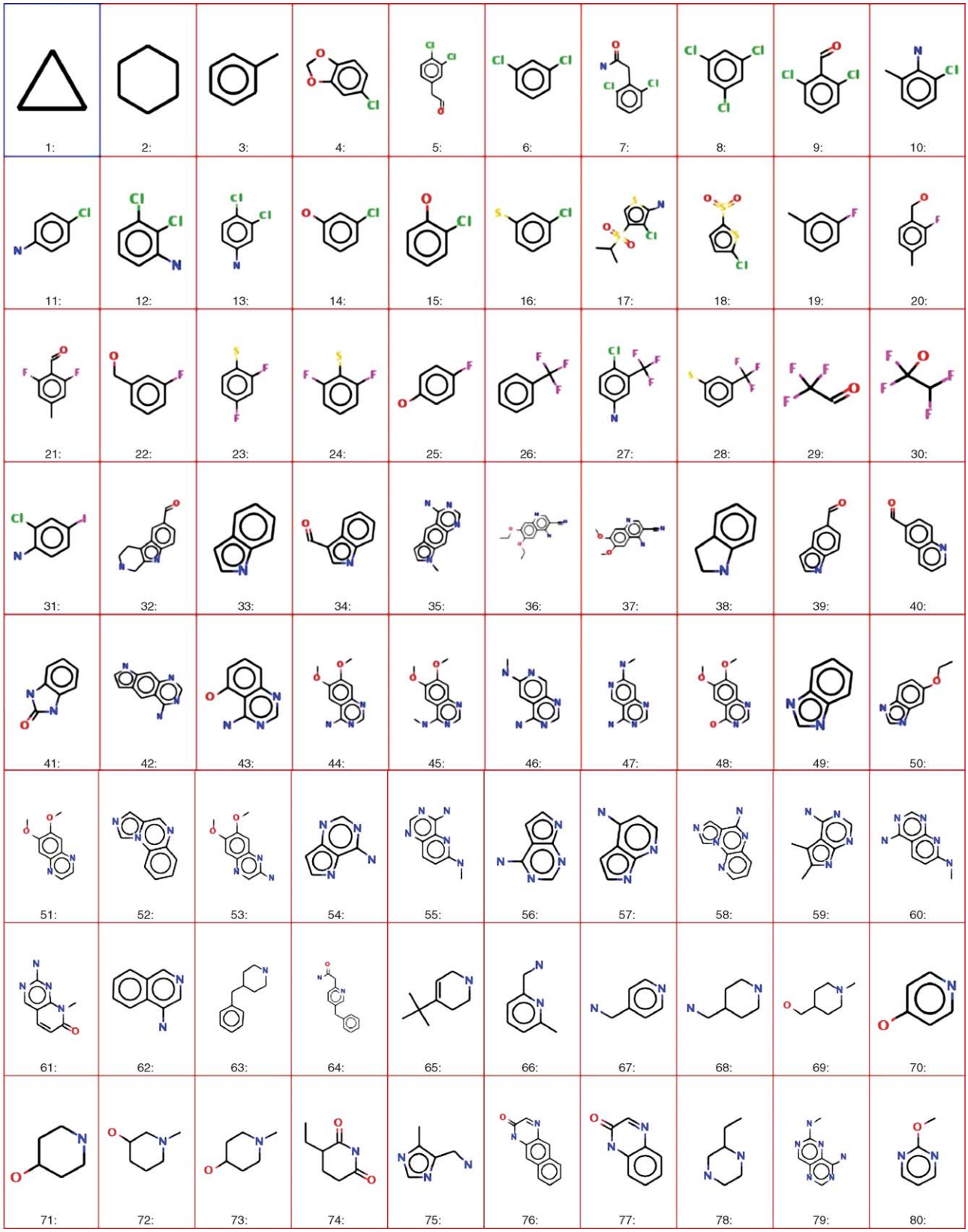


Fig. 2. (*Continued*).

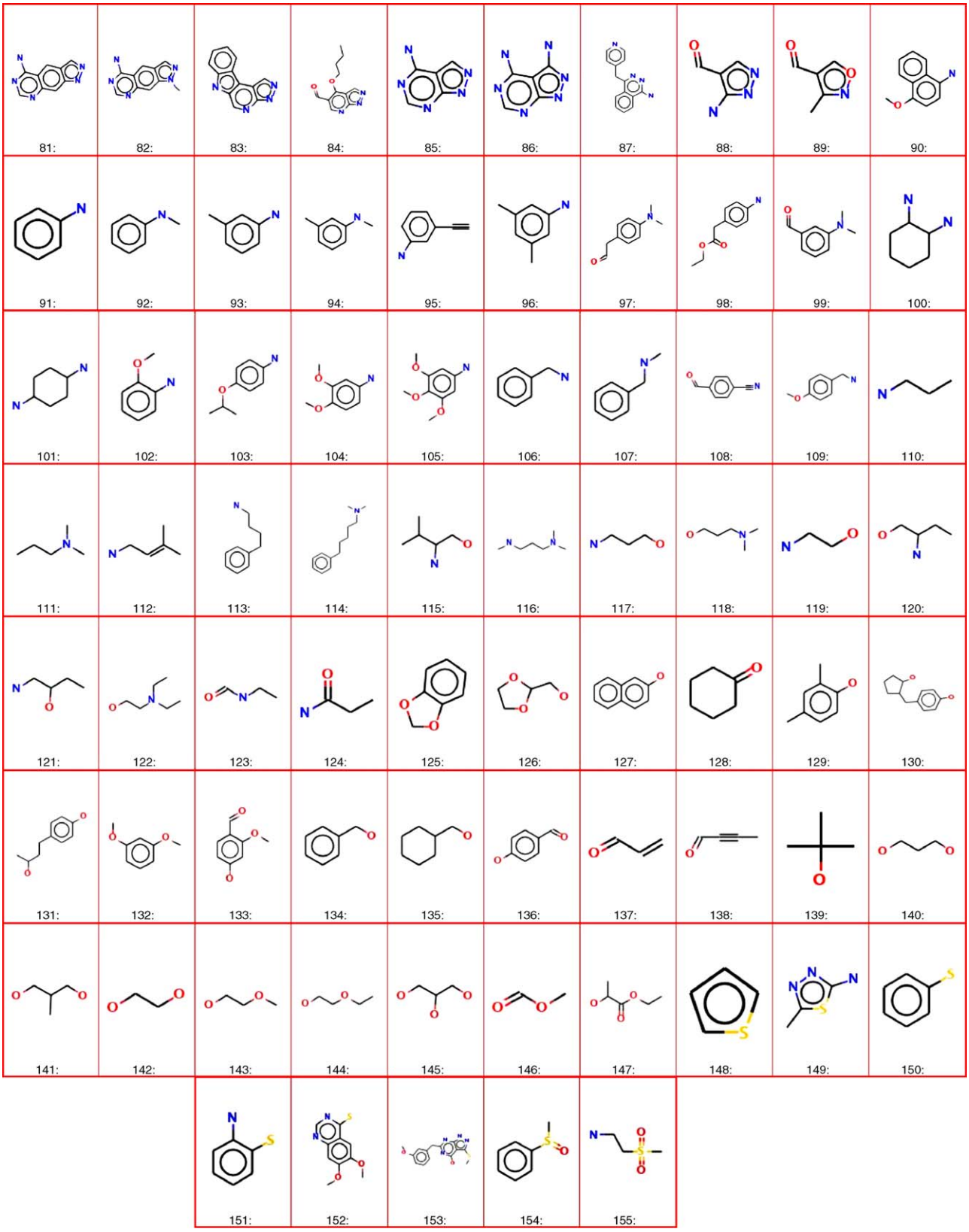
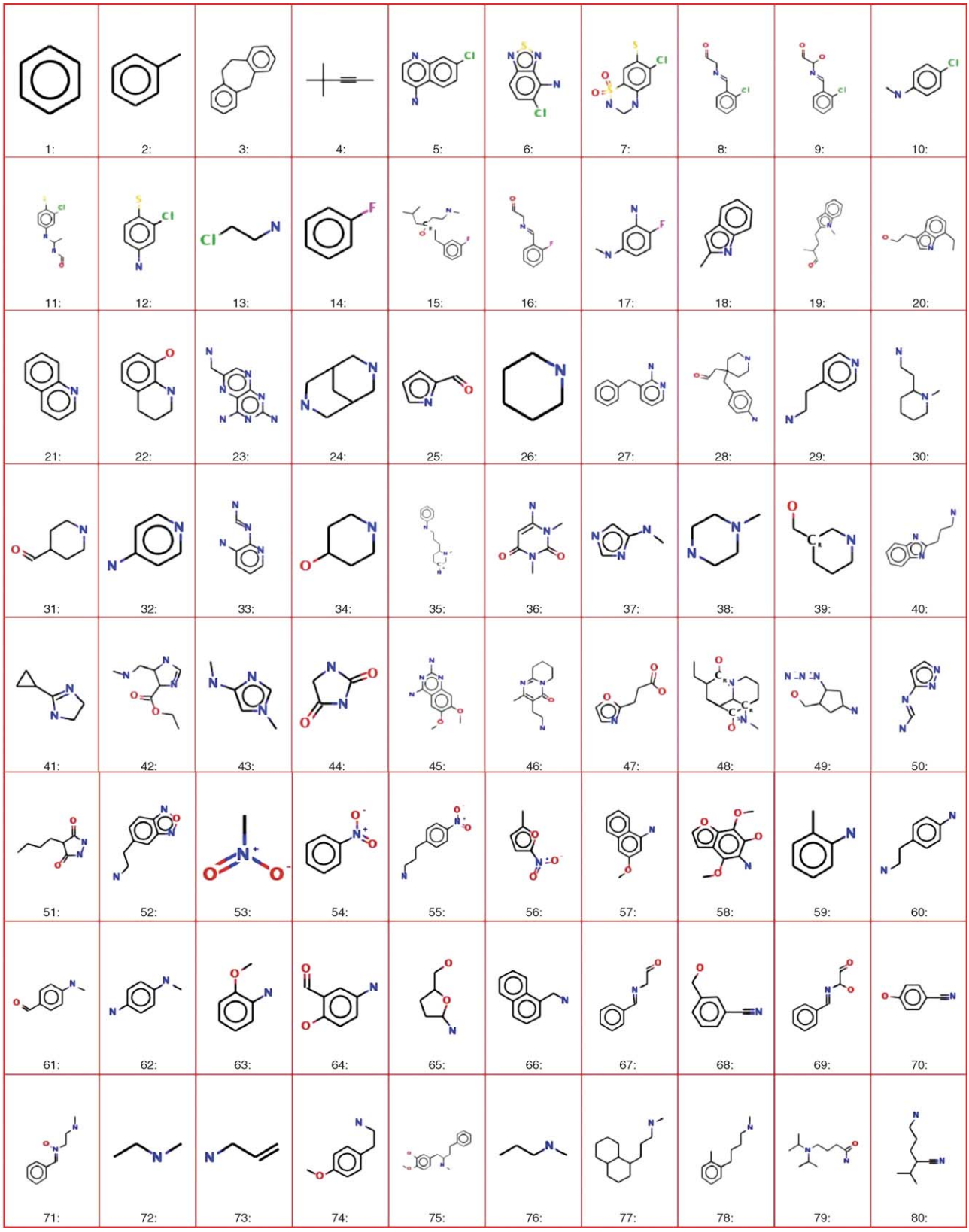


Fig. 2. (*Continued*).



(A)

Fig. 3. (A) Scaffolds and (B) sidechains found in the Yoshida–Topliss database but not found in the kinase dataset.

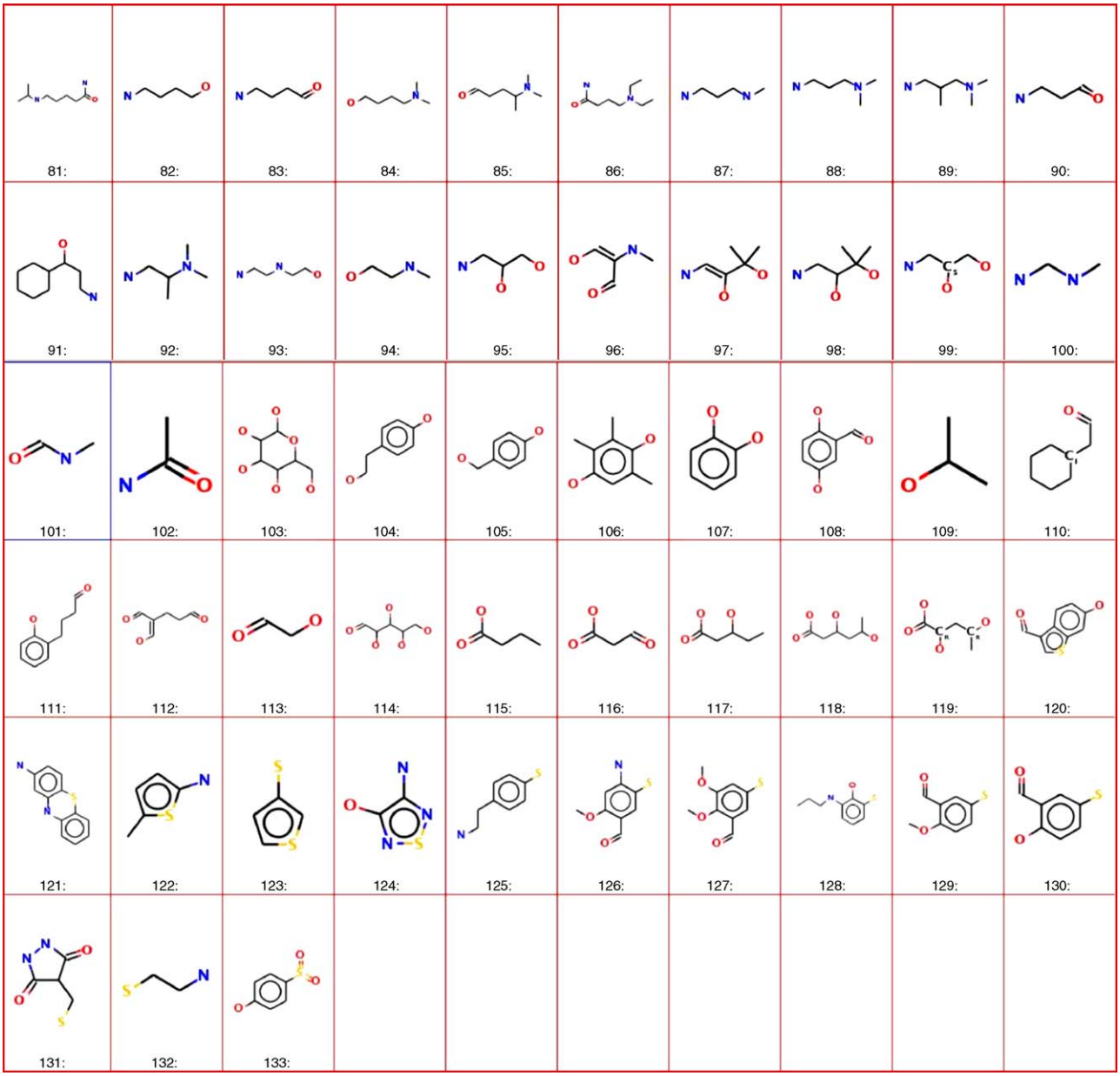


Fig. 3. (*Continued*).

recognition is on average 88% and at worse 86%. The false positive rate is on average 9% and at worse 11%. This recognition behavior is only slightly inferior to the performance of the whole dataset training case applied to itself. Training over 20% would seem on average to be little deteriorated from inspection of the average recognition of 82% but the WCS value is 75% signifying that we have moved into a less consistent performance area. The last partition presented is the case of a 10% training set and by this point the WCS recognition value is 69% and the average is 80%. False positives for both the 20 and 10% trails are relatively low. Below 10%, the absolute populations for the training sets are below 50 members randomly selected from the combined >500 member union of the kinase dataset and Yoshida–Topliss dataset. These trials showed large gaps between averages and WCS values and the false positive rate moved towards 30%. While this is still better than random, it is not desirable performance. It is however, somewhat expected since it is difficult to develop adequate diversity with only dozens of compounds.

The employed metrics in the model are all 2D or 1D metrics so there are no conformational sensitivity issues or determining the active conformation problem. It is a simple screen to determine a Boolean quality to aid decision making in compound selection for high throughput screening and performs well with the internal reserve and test strategy outlined above. Why do these descriptors work? The answer is simply that kinase inhibitors are in a distinct property space versus presently marketed orally available drug molecules. This can be seen by the distributions presented in Fig. 5.

(B)

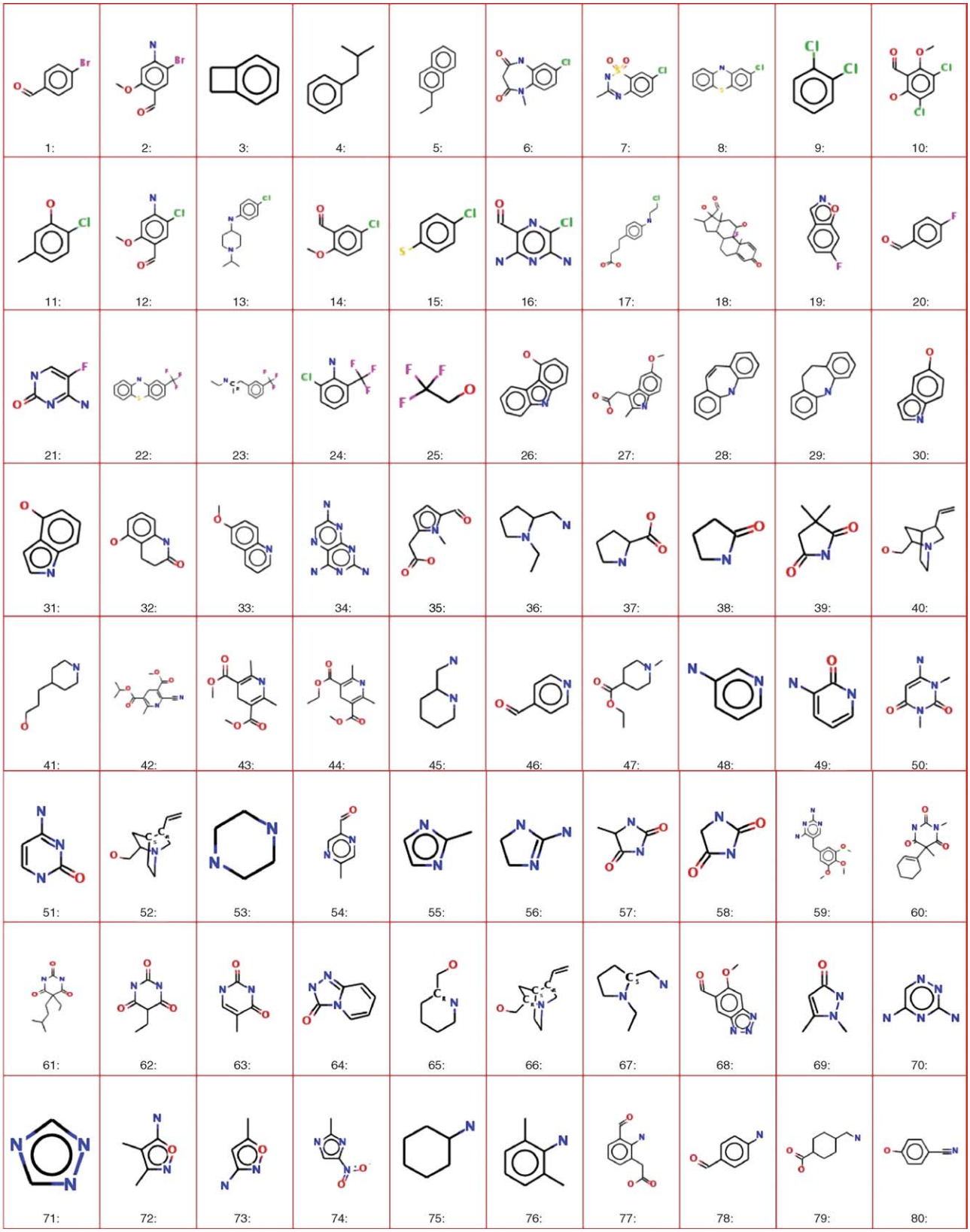


Fig. 3. (*Continued*).

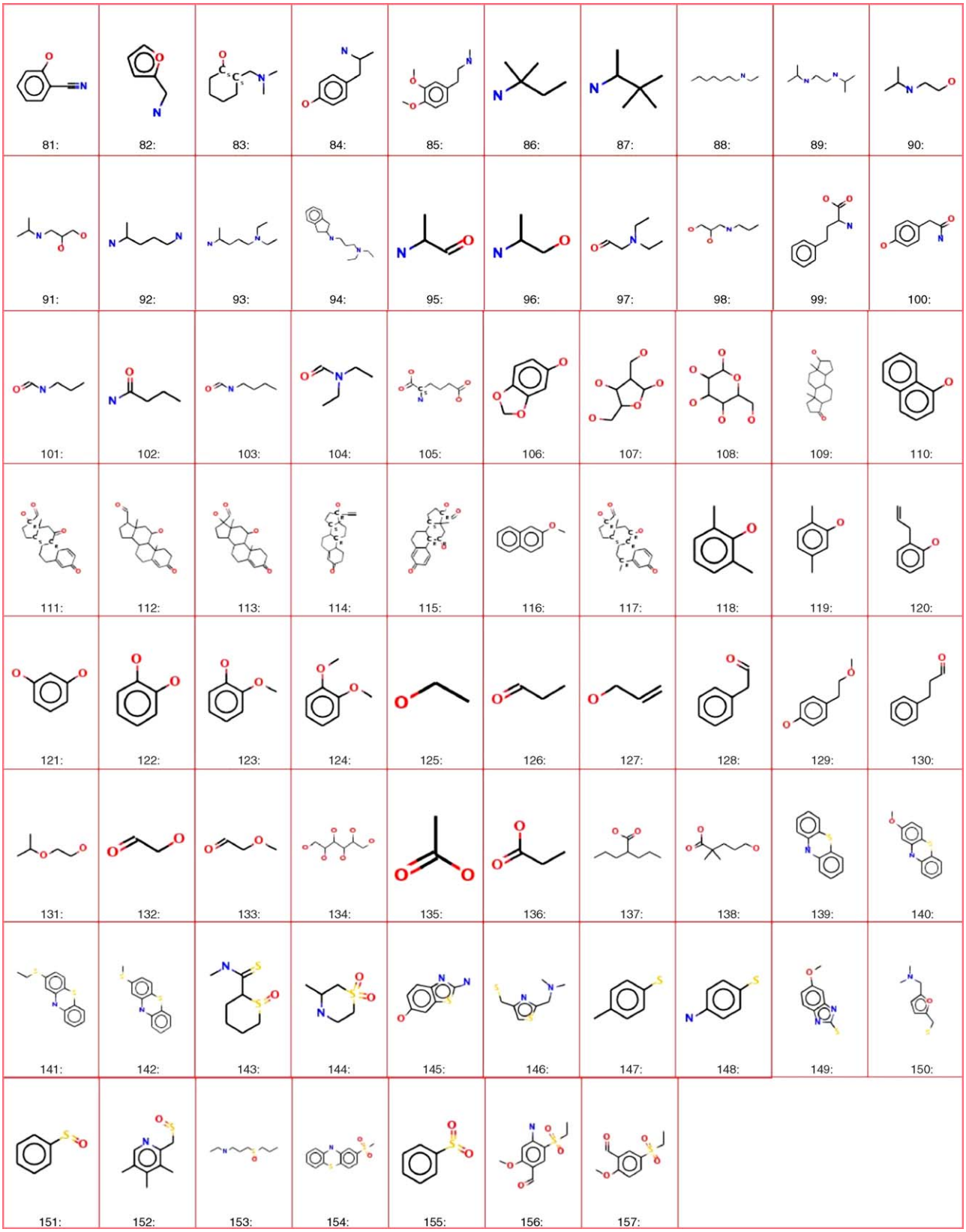


Fig. 3. (*Continued*).

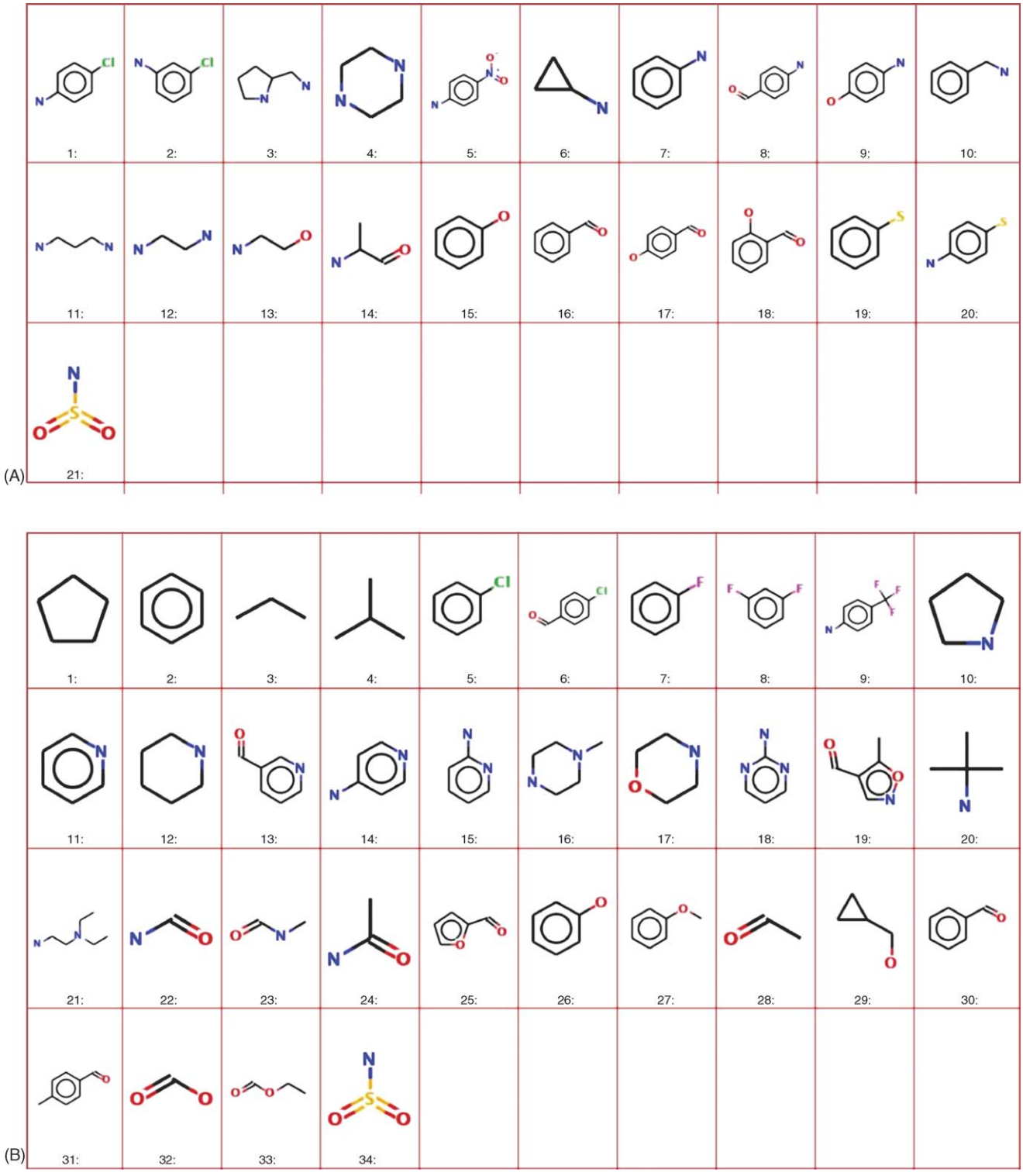


Fig. 4. (A) Scaffolds and (B) sidechains common to both the Yoshida–Topliss and kinase datasets.

In Fig. 5A–H, in distribution after distribution, a distinct difference can be seen between the Yoshida–Topliss and kinase inhibitor datasets. Fig. 5A illustrates a distinct greater number of aromatic carbons in kinase inhibitors. This matches intuition, which many medicinal chemist have that kinase inhibitors are relatively flat as to match the shape of the kinase cleft itself. This aromatic content is more complicated than simple carbon and is indication for other traits, which are seen in additional figures. The majority of the Yoshida–Topliss dataset does not have aromatic nitrogens (Fig. 5B) while the converse is true for kinase inhibitors. Further, all kinase inhibitors have at least two ring systems. Five member rings are rare in the Yoshida–Topliss

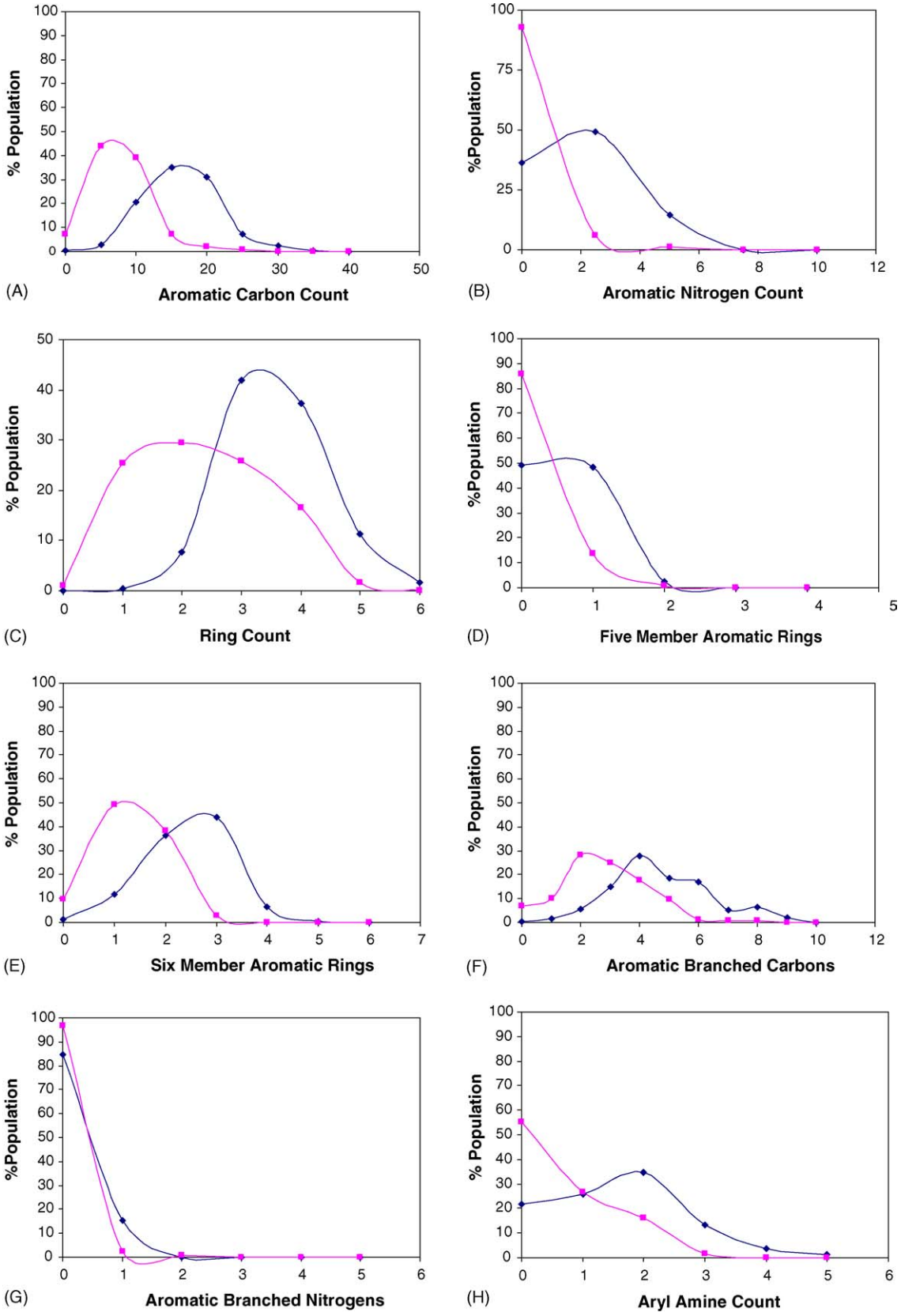


Fig. 5. Distribution comparison between the Yoshida–Topliss and kinase datasets for various descriptors. In all graphs, % population refers to the local population at a different value of the specific descriptor of interest, pink squares are for the Yoshida–Topliss dataset and dark blue diamonds denote the kinase dataset. (A) Aromatic carbon count. (B) Aromatic nitrogen count. (C) Ring count (aromatic or saturated). (D) Five member aromatic ring count. (E) Six member aromatic ring count. (F) Aromatic branched carbon count. (G) Aromatic branched nitrogen count. (H) Aryl amine count.

Table 4
Prediction of the kinase content for various vendor datasets

Kinase datasets	
Vendor	% Identified
Asinex	73
BioFocus	89
ChemDiv	66
LifeChemicals	40

dataset while about half of the kinase dataset have one while some have two. Multiple six member aromatic rings are essentially required for kinase inhibitors from a statistical standpoint. Heterocyclic structures in kinase inhibitors are simply more complex as is seen by more branched aromatic carbons and nitrogens (Fig. 5F and G). The difference in Fig. 5G between kinase and the Yoshida–Topliss dataset may appear slight at first glance but it is a difference of 15% versus 3% having branched aromatic nitrogens. A last presented metric is the aryl nitrogen count, which is common in both but more prominent in kinase inhibitors. Not shown were metrics that favored Yoshida–Topliss, which included flex and saturated atom counts. Molecular weight, atom and bond counts, were seen to have no differences in distributions between the two datasets. For this reason, we removed these descriptors from our final models and saw no detriment. These metrics present an image of kinase inhibitors as a population being more rigid and more heterocyclically complex that currently marketed pharmaceuticals.

The motive for all this work was to have an independent and quick screen for assessing what compound libraries should be purchased for screening against our kinase targets. Being a small company, we are not able to purchase all of the libraries. Since we have novel targets, our proteins are unique to us and stocks are a capital resource. In this section, we describe our findings of our application to analysis for some common commercial kinase and nonkinase libraries. The list in Table 4 is short since we initially confined ourselves to vendors with relatively large and relatively diverse kinase datasets. Diversity was evaluated as the number of distinct clusters determined at the value of 0.85 for the Tanimoto coefficient with MACCS 2D fingerprints. A compound was said to be recognized if the QSAR model assigned a value of −1.00 to 0.5 to the compound. The best performance overall was seen by the BioFocus datasets which had a recognition rate of 89%. Asinex and ChemDiv both were recognition rates near 70%. LifeChemicals was a poor performer with a low recognition rate but it is a large dataset so it is possible to select a well performing subset easily. Vendor datasets advertised as being for generic screening we seen to have lower recognition rates (20–30%) and had more compounds that were simply not within the use radius of the model (i.e., the model assigned values outside the −1.00 to 0.5 range).

With this recognition function and with considerations of diversity, we made our purchases. Since BioFocus had the superior set of compounds, we purchased from them first, with a requirement for recognition. To maintain diversity, we selected a spread across the available scaffolds seen in the BioFocus datasets. In additional purchases from Asinex, scaffolds previously purchased from BioFocus were eliminated from consideration then passed though the screen. Additional diversity selections were performed with OptiSim [36]. Last purchases were made from LifeChemicals, which has a large population from which to select.

Though vendors have been reluctant to display exactly how they design and validate their kinase focus libraries, tools like the QSAR model outlined here are easily assembled once the recognition of the proper descriptors that are derived. Alternatively to a QSAR approach, others have developed machine learning and artificial intelligence based methods [37].

4. Summary

Herein, we have described in this paper a simple QSAR model able to recognize kinase inhibitors from other druglike compounds. It is heuristic in that it is based on 258 known potent kinase inhibitors and 230 marketed pharmaceuticals. The dataset is itself an important factor of the work: the training is done with potent inhibitors and marketed drugs. The method is based on 2D and 1D descriptors and so does not have conformational sensitivity and is fast compared to docking or to 3D pharmacophore based methods. The reasons for the model working can be seen from the population distributions seen in Fig. 5: the descriptors allow for the definition of a kinase-inhibitor space, which is distinct from the space occupied by presently marketed orally available drugs. This opens the obvious possibility for the employment of these descriptors as filters to enrich HTS compound datasets with more kinase-inhibitor like compounds. The protocol is able to be trained over small portions of our datasets and still have kinase-inhibitor recognition rates >85% and false positive rates below 10%. The methodology shows distinct trends differences between libraries advertised as being kinase focus versus those advertised as being for generic HTS. Further, when applied to different vendors, it provides an independent means of evaluating the kinase inhibitor likeness quality of each dataset to facilitate purchase decisions.

The use of models for descrimination between target class specific molecules is its early phases, which contrast with the abundance of studies that map the space of druglike molecules [20,21,30–33]. The pioneers in this have been library vendors who published little, probably to protect their intellectual property. Recent papers using artificial intelligence (AI) and neural networks (NN) [37,38] have, like the present study, demonstrated that kinase inhibitors are distinct from other compounds. The present study is unique in developing a QSAR model (rather than applying NN or AI) based on meaningful descriptors determined from a systematic comparison of the fragment space available to both a kinase inhibitor dataset and a dataset of marketed drugs. It is likely that additional studies in this genre will map the distinct property space that kinase inhibitors occupy and stimulate more rapid discovery of therapeutics.

Acknowledgment and additional data

The authors graciously thank Dr. J. Swanson (Tripos Inc.) for providing a copy of his implementation of the RECAP protocol. Copies of the database used for training this QSAR model and an SPL script for generating the descriptors is available from dsprous@cytrxlabs.com.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2005.09.004.

References

[1] S.K. Hanks, T. Hunter, Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification, FASEB 9 (1995) 576–596.
[2] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, The protein kinase complement of the human genome, Science 298 (2002) 1912–1934.
[3] N.G. Ahn, K.A. Resing, Cell biology. Lessons in rational drug design for protein kinases, Science 308 (2005) 1266–1267.
[4] M.S. Cohen, C. Zhang, K.M. Shokat, J. Taunton, Structural bioinformatics-based design of selective, irreversible kinase inhibitors, Science 308 (2005) 1318–1321.
[5] H. Weinmann, R. Metternich, Drug discovery process for kinase inhibitors, Chem. Biol. Chem. 6 (2005) 455–459.
[6] O. Prien, Target-family-oriented focused libraries for kinases—conceptual design aspects and commercial availability, Chem. Biol. Chem. 6 (2005) 500–505.
[7] B.J. Druker, S. Tamura, E. Buchdunger, S. Ohno, G.M. Segal, S. Fanning, J. Zimmermann, N.B. Lydon, Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells, Nat. Med. 2 (1996) 561–569.
[8] A.J. Barker, K.J. Gibson, W. Grundy, A.A. Godfrey, J.J. Barlow, M.P. Healy, J.J. Woodburn, S.E. Ashton, B.J. Curry, L. Scarlett, L. Henthorn, L. Richards, Studies leading to the identification of ZD1839 (IRESSA): an orally active, selective epidermal growth factor receptor tyrosine kinase inhibitor targeted to the treatment of cancer, Bioorg. Med. Chem. Lett. 11 (2001) 1911–1914.
[9] S.G. O'Brien, F. Guilhot, R.A. Larson, I. Gathmann, M. Baccarani, F. Cervantes, J.J. Cornelissen, T. Fischer, A. Hochhaus, T. Hughes, K. Lechner, J.L. Neilsen, P. Rouselot, J. Reiffers, G. Saglio, J. Shepherd, B. Simonsson, A. Gratwohl, J.M. Goldman, H. Kantarjian, K. Taylor, G. Verhoef, A.E. Bolton, R. Capdeville, B.J. Druker, Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia, N. Engl. J. Med. 348 (2003) 1048–1050.
[10] J.S. Ochs, Rationale and clinical basis for combining gefitinib (IRESSA, ZD1839) with radiation therapy for solid tumors, Int. J. Radiat. Oncol. Biol. Phys. 58 (2004) 941–949.
[11] T. Asano, I. Ikegaki, S. Satoh, M. Seto, Y. Sasaki, A protein kinase inhibitor, Fasudil (AT-877): A novel approach to signal transduction therapy, Cardiovasc. Drug Rev. 16 (1998) 76–78.
[12] S.R. Whittaker, M.I. Walton, M.D. Garrett, P. Workman, The cyclin-dependent kinase inhibitor CYC202 (*R*-Roscovitine) inhibits retinoblastoma protein phosphorylation, causes loss of Cyclin D1, and activates the mitogen-activated protein kinase pathway, Cancer Res. 64 (2004) 262–272;
L.A. Sobera, J. Castaner, J. Bozzo, P.A. Leeson, Drugs Future 27, 1141–1147.
[13] J. Drevs, M. Medinger, C. Schmidt-Gersbach, R. Weber, C. Unger, Receptor tyrosine kinases: The main targets for new anticancer therapy, Curr. Drug Targets 4 (2003) 113–121.
[14] C.J. Sawyers, Opportunities and challenges in the development of kinase inhibitor therapy for cancer, Genes Dev. 17 (2003) 2998–3010.
[15] M. Cherry, D.H. Williams, Recent kinase and kinase inhibitor X-ray structures: Mechanisms of inhibition and selectivity insights, Curr. Med. Chem. 11 (2004) 663–673.
[16] T. Naumann, H. Matter, Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes, J. Med. Chem. 45 (2002) 2366–2378.
[17] J.R. Woolfrey, G.S. Weston, The use of computational methods in the discovery and design of kinase inhibitors, Curr. Pharm. Des. 8 (2002) 1527–1545.
[18] P. Traxler, P. Furet, Strategies toward the design of novel and selective protein tyrosine kinase inhibitors, Pharmacol. Ther. 82 (1999) 195–206.
[19] WO03004147; WO9926901.
[20] G. McMahon, L. Sun, C. Liang, C. Tang, Protein kinase inhibitors: structural determinants for target specificity, Curr. Opin. Drug Disc. Dev. 1 (1998) 131–146;
A. Bridges, Chemical inhibitors of protein kinases, Chem. Rev. 101 (2001) 2541–2571;
J.L. Adams, D. Lee, Recent progress towards the identification of selective inhibitors of serine/threonine protein kinases, Curr. Opin. Drug Disc. Dev. 2 (1999) 96–109.
[21] F. Yoshida, J.G. Topliss, QSAR model for drug human oral bioavailability, J. Med. Chem. 43 (2000) 2575–2585.
[22] T.I. Oprea, J. Gottfries, A one component model for oral bioavailability, J. Mol. Graph. Model. 5 (1999) 261–274.
[23] X.Q. Lewell, D.B. Judd, S.P. Watson, M.M. Hann, RECAP, J. Chem. Inf. Comput. Sci. 38 (1998) 511–522.
[24] SYBYL 6.9.1. 2003. Tripos Inc., St. Louis MO 63144, USA.
[25] Asinex: www.asinex.com.
[26] BioFocus: www.biofocus.com.
[27] ChemBridge: www.chembridge.com.
[28] ChemDiv: www.chemdiv.com.
[29] LifesChemicals: www.lifeschemicals.com.
[30] P. Ertle, B. Rohde, P. Selzer, fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, J. Med. Chem. 43 (2000) 3714–3717.
[31] J.F. Blake, Chemoinformatics—predicting the physicochemical properties of 'drug-like' molecules, Cur. Opin. Biotechnol. 11 (2000) 104–107.
[32] J.F. Blake, Examination of the computed molecular properties of compounds selected for clinical development, BioTechniques 34 (2003) 16–20.
[33] H. van de Waterbeemd, G. Camenisch, G. Folker, J.R. Chretien, O.A. Raevsky, Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors, J. Drug Target. 15 (1998) 480–490.
[34] S. Wold, E. Johansson, M. Cocchi, PLS, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, ESCOM, Leiden, Germany, 1993, pp. 523–550.
[35] R.D. Clark, D.G. Sprous, J.M. Leonard, Progressive scrambling, in: H.-D. Holtje, W. Sippl (Eds.), Rational Approaches to Drug Design, Prous Science, S.A. Barcelona, Spain, 2001, pp. 475–486.
[36] R.D. Clark, Optisim, J. Chem. Inf. Comput. Sci. 37 (1996) 1181–1188.
[37] M.G. Ford, W.R. Pitt, D.C. Whitley, Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks, J. Mol. Mod. Graph. 22 (2004) 464–467.
[38] H. Briem, J. Gunther, Classifying "kinase inhibitor likeness" using machine learning methods, Chem. Biol. Chem. 6 (2005) 558–566.