

Modeling and Benchmark Data Set for the Inhibition of c-Jun N-terminal Kinase-3

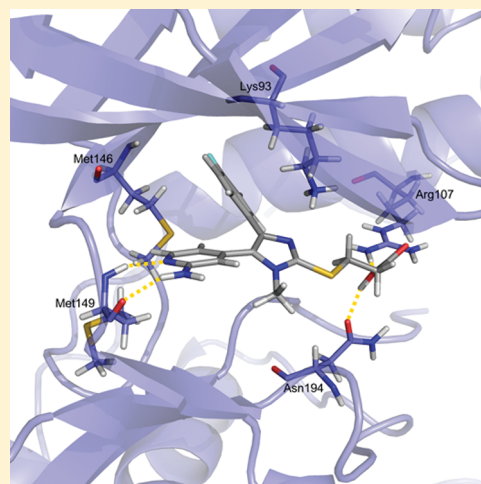
Verena Schattel,[†] Georg Hinselmann,[‡] Andreas Jahn,[‡] Andreas Zell,[‡] and Stefan Laufer^{*,†}

[†]Department of Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Eberhard Karls University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany

[‡]Center for Bioinformatics (ZBIT), Eberhard Karls University of Tübingen, Sand 1, 72076 Tübingen, Germany

 Supporting Information

ABSTRACT: The goal of this paper is to present and describe a novel 2D- and 3D-QSAR (quantitative structure–activity relationship) binary classification data set for the inhibition of c-Jun N-terminal kinase-3 with previously unpublished activities for a diverse set of compounds. JNK3 is an important pharmaceutical target because it is involved in many neurological disorders. Accordingly, the development of JNK3 inhibitors has gained increasing interest. 2D and 3D versions of the data set were used, consisting of 313 (70 actives) and 249 (60 actives) compounds, respectively. All compounds, for which activity was only determined for the racemate, were removed from the 3D data set. We investigated the diversity of the data sets by an agglomerative clustering with feature trees and show that the data set contains several different scaffolds. Furthermore, we show that the benchmarks can be tackled with standard supervised learning algorithms with a convincing performance. For the 2D problem, a random decision forest classifier achieves a Matthew's correlation coefficient of 0.744, the 3D problem could be modeled with a Matthew's correlation coefficient of 0.524 with 3D pharmacophores and a support vector machine. The performance of both data sets was evaluated within a nested 10-fold cross-validation. We therefore suggest that the data set is a reasonable basis for generating QSAR models for JNK3 because of its diverse composition and the performance of the classifiers presented in this study.



INTRODUCTION

The aim of this study is to present and describe a novel 2D- and 3D-QSAR data set for c-Jun N-terminal kinase-3 (JNK3) inhibition as a binary classification problem.

JNK3 is a member of the mitogen activated protein kinases (MAPKs) and is involved in the modulation of neuronal apoptosis. Dysfunctions of JNK3 lead to diseases like Parkinson's disease, Alzheimer's disease, multiple sclerosis, epilepsy, stroke, and other dysfunctions of the central nervous system. Localized expression of the JNK3 in the brain makes it an interesting drug target.^{1,2} The close homology within the ATP-binding site of more than 90% sequence identity of the JNK isoforms and the close homology to other MAPK members, like the p38 α or the extracellular-signal regulated kinase ERK2, complicates the design of selective and isoform specific inhibitors.²

Other authors have investigated JNK3 inhibitors in combination with QSAR approaches. Shaikh et al.,³ for example, presented a 3D-QSAR approach in combination with docking studies on 44 (benzothiazole-2-yl)acetonitrile derivatives as JNK3 inhibitors in their work yet without publishing the used data set.³ A binary 2D-QSAR approach was conducted on the proprietary Aureus Kinase knowledge database by Ijjaali et al.⁴

Chung et al. investigated a receptor-guided 3D-QSAR analysis of anilinothiopyridines as JNK3 inhibitors.⁵ They used the methodologies of CoMFA (comparative molecular field analysis) and CoMSIA (comparative molecular similarity indices analysis) for their studies.

In our data sets, we only used compounds, which were developed and synthesized in our group with activities measured all in the same assay. Some of these compounds were originally developed as putative p38 α inhibitors. Nevertheless, several of the compounds inhibit JNK3 MAPK. The classes of compounds included in the data set are methylsulfanyl imidazoles,⁶ benzylsulfanyl imidazoles,^{6,7} substituted isoxazole,⁸ N-substituted 5-isoxazolones,⁹ acetylaminopyridines,¹⁰ [4-(methylsulfanyl)benzyl]-sulfanyl-substituted imidazole compounds,¹⁰ morpholine-substituted imidazoles,¹⁰ and diverse purine derivatives.¹¹ The ligands were prepared with Schrödinger LigPrep.¹² Compounds with stereogenic centers were discarded from the 3D data set, unless both, the R- and S-isomer had been determined. Thus, we can provide a clean and unpublished 2D and 3D data set of JNK3 inhibitors and inactive analogues.

Received: October 19, 2010

Published: January 31, 2011

To analyze the diversity of the data sets, we applied agglomerative clustering with Feature Trees with different distance cutoffs. The results of the clustering are included in the SD files as property tags.

We set up QSAR modeling approaches to ensure that the resulting binary classification problem can be modeled with a reasonable quality. A major goal of this study was to investigate whether the actives and inactives can be separated by state-of-the-art machine learning techniques. The basic machine learning algorithms and encodings are well-established in the field of cheminformatics. Similar approaches can be found, for instance, in the works of Svetnik et al.,¹³ Bender et al.,¹⁴ and Hsieh et al.¹⁵ The results show that the problem could be learned using various molecular encodings with straightforward machine learning approaches. The results for the 2D benchmark reach up to an average Matthew's correlation coefficient (MCC) of 0.744 using a random decision forest classifier. The 3D benchmark was modeled with a 3D pharmacophore encoding and a support vector machine yielding an average MCC of 0.524. These results can be considered as reference results for further research. Our results of the machine learning approaches show that the data set is a reasonable basis for generating QSAR models for JNK3.

To summarize, we think that the data set presented in this work is a valuable contribution to the research field because the affinities were determined in the same assay, the diversity analysis indicates a high structural diversity, and the QSAR problem could be learned with convincing results.

MATERIALS AND METHODS

In this section, we first explain the pharmaceutical background of the provided data set, which includes the description of the target as well as the characterization of the several structure classes and their preparation protocol. Afterward, a detailed description of the diversity analysis and the molecular encodings and machine learning algorithms follows.

Biological Background. The c-Jun N-terminal kinases, with their isoforms JNK1, JNK2, and JNK3, belong to the family of mitogen activated protein kinases. The JNK subfamily consists of ten known isoforms encoded by three genes, *jnk-1*, *jnk-2*, and *jnk-3*. This work relates to the JNK3, which is expressed primarily in the brain and at low levels in the testes and kidney. JNK1 and JNK2 are ubiquitously expressed.¹⁶

Besides the JNK family, the MAPKs include the p38 MAPKs with the four isoforms p38 α , p38 β , p38 γ , and p38 δ as well as the extracellular-signal regulated kinases ERK1, ERK2, ERK3/4, ERK5, and ERK7/8. All MAPKs are serine/threonine protein kinases belonging to a number of 518 different protein kinases, which thus represent the largest family of genes in eukaryotes. Protein kinases mediate most of the signal transduction by modification of substrate activity and also control many other cellular processes, including metabolism, transcription, cell cycle, progression, apoptosis, and differentiation.¹⁷ Especially the JNK3 has been shown to modulate neuronal apoptosis and may be involved in the pathology of neurodegenerative diseases. The inhibition of JNK3 could be useful for treating Parkinson's disease, epilepsy, stroke, and other dysfunctions of the central nervous system.¹ Accordingly, the development of JNK3 inhibitors has increasingly gained interest.

The binding site of the kinases (Figure 1), in which the natural cosubstrate ATP binds, is located in the center of the enzyme between the N-terminal and C-terminal domain. The two lobes

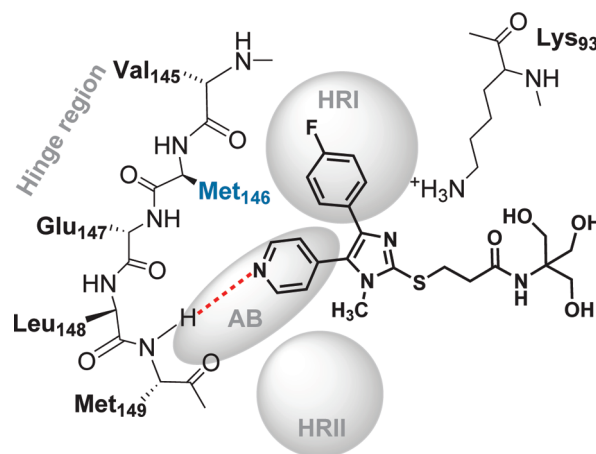


Figure 1. Schematic binding mode of the active compound **49** in the JNK3 binding site. The hydrogen bonds are shown as red dotted lines. The gatekeeper residue Met146 is shown in blue and the hydrophobic regions I (HRI) and II (HRII) as well as the adenine binding region (AB) are sketched in gray. The possible binding mode of the compound was obtained by docking studies with the induced fit docking protocol from Schrödinger²⁰ on the X-ray structure PDB 1PMQ.¹⁹ The OH-groups are able to build hydrogen bonds with the surrounding amino acids. However, for clarity these interactions are not shown in this figure.

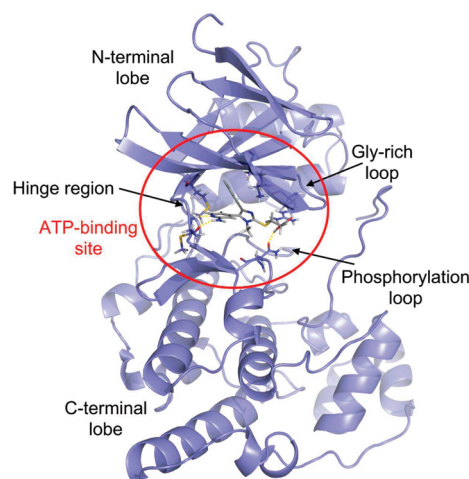
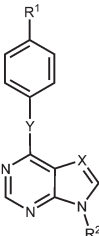
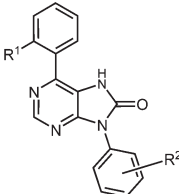


Figure 2. The structure of a JNK3 MAPK (PDB code 1PMQ¹⁹) with the active compound **49** docked into the ATP-binding site.

are connected by the so-called hinge region which contains amino acids (Met149 in JNK3) that interact with the bound ATP. This ATP-binding site is conserved in all kinases and contains, according to Traxler et al.,¹⁸ five diverse regions: (1) the hydrophobic adenine binding region, (2) the hydrophobic region I (also called selectivity pocket), (3) the hydrophilic sugar pocket, (4) the phosphate binding region, and (5) the hydrophobic region II, which is located in the front of the binding site and therefore exposed to solvent. The two hydrophobic regions I and II are not occupied by bound ATP but do contain residues that vary among kinases. Because of this, the residue variance in those pockets can be used for the design of potent and selective inhibitors.¹⁹ Due to the structure homology within the MAPKs, it is a challenging task to design potent and selective JNK3 inhibitors.

Table 1. Purine Derivatives as Potential ATP-Competitive Kinase Inhibitors^{11,26}



#	X	Y	R ¹	R ²	Inhibition
02	-N	-NH	-H	-H	0
03	-N	-NH	-F	-H	0
04	-N	-SCH ₂	-H	-H	0
05	-N	-SCH ₂	-F	-H	0
06	-N	-S	-H	-H	0
07	-N	-S	-F	-H	0
10	-C	-S	-H	-H	0
11	-C	-S	-F	-H	0

#	R ¹	R ²	Inhibition
14	-Cl	2,6-di-F	0
15	-H	2,6-di-F	1
16	-Me	2,6-di-F	0
17	-Cl	2,4-di-F	0
18	-Me	2,4-di-F	0
19	-Me	2,4-di-OMe	0
20	-F	2,6-di-F	0

The data set presented in this work contains 313 (2D) or 249 (3D) compounds that were investigated for their ability to inhibit JNK3. All ligands were synthesized in the group of Prof. Laufer at the University of Tübingen. They are all previously unpublished with respect to the inhibition of the JNK3 MAPK. All ligands are ATP-competitive and therefore bind in the ATP binding site of the JNK3 MAPK. The data set contains mainly derivatives with a five-membered ring as the core of the scaffold. Furthermore, the data set comprises methylsulfanyl imidazoles,⁶ benzylsulfanyl imidazoles,^{6,7} substituted isoxazole,⁸ N-substituted 5-isoxazolones,⁹ acetylamino-pyridines,¹⁰ [4-(methylsulfanyl)benzyl]sulfanyl-substituted imidazole compounds,¹⁰ morpholine-substituted imidazoles,¹⁰ and diverse purine derivatives.¹¹

Purine Derivatives as Potential ATP-Competitive Kinase Inhibitors. Purine derivatives have already been reported as kinase inhibitors, mainly as inhibitors for cyclin-dependent kinases (CDKs). One of those substituted purines is for example Olomoucine, a 2,6,9-substituted purine which inhibits CDK2 with an IC₅₀ value of 7 μ M.²¹ In our design of purine derivatives, we combined the purine system from the cosubstrate ATP and phenyl moieties. With those kinds of compounds we wanted to explore possible interactions with different regions of the ATP binding site in disease-related protein kinases, such as the JNK3 MAPK. An important difference to CDK2 is that the hydrophobic region I in JNK3 is larger. To address this increased hydrophobic region, phenyl moieties were introduced. In order to exploit all possible orientations of the pyrine system according to the binding properties of the phenyl moieties in the hydrophobic region, various linkers for the introduction of those phenyl moieties were used.¹¹ A series of about 20 purine derivatives was synthesized in our group and tested for inhibitory potency of JNK3 MAPK. Unfortunately, only one compound of the 20 tested inhibits JNK3 potently, as shown in Table 1.

Isoxazolone Based Compounds. Isoxazolone based inhibitors were designed by transfer of SAR studies from N3-substituted imidazoles to isoxazolones. The isoxazole was used as a bioisosteric

replacement of the imidazole heterocycle.⁹ There are 23 isoxazolone based inhibitors in the data set with 9 active compounds.

Substituted 4,5-Diarylisoxazoles. This class of compounds was created by bioisosteric replacement of the imidazole ring from SB203580 to an isoxazole ring. SB203580 is cocrystallized with p38 α MAPK in the X-ray structure with the PDB code 1A9U.²² SB203580 also inhibits the JNK3 MAPK.²³ 2,3,4- and 4,5-disubstituted as well as 3,4,5-trisubstituted isoxazole derivatives are used in this data set. Five of these compounds show JNK3 inhibition.

Diverse Imidazole Inhibitors. Polysubstituted pyridin-4-yl imidazole inhibitors of MAPKs were prepared as small molecular anticytokine agents. Additionally, these inhibitors are drug candidates for the treatment of chronic inflammatory diseases.⁶ As all the other compound classes described here, some of the presented pyridinyl imidazoles are potent p38 α inhibitors as well. There are several pyridinyl imidazoles which inhibit the JNK3 MAPK. The 2D-data set contains 68 pyridinyl imidazoles. Ten compounds were labeled as active JNK3 inhibitors. Furthermore, there are 59 2,4,5-trisubstituted imidazole derivatives that contain methylsulfanyl and benzylsulfanyl imidazoles. Eighteen of those are JNK3 inhibitors. Another group of compounds consists of 1,2,4,5-tetrasubstituted imidazole derivatives. Originally, they were synthesized to demonstrate the effects of substitution at imidazole N1 and were designed to inhibit the p38 α MAPK.¹⁰ There are 114 tetrasubstituted imidazoles in the data set. Twenty-seven compounds of them are JNK3 inhibitors.

Data Set Preparation Protocol. The structures of a variety of compounds from the data set are listed in the Tables 1-5. The compound number of each structure, presented in the tables, matches the number, under the tag 'name', of the respective structure in the MDL SD file of the 2D data set. The low-energy conformations of the 2D and 3D structures were optimized with LigPrep¹² version 2.2 from the Schrödinger Software Suite²⁴ version 8.5. After preparation with LigPrep we annotated the

Table 2. Isoxazolone Based Compounds⁹ as Inhibitors of JNK3

#	R ¹	Inhibition	#	R ¹	Inhibition
144	-CH ₂ -CH ₃	0	151		0
145		1	152	-(CH ₂) ₂ -OH	0
146	-CH ₂ -O-CH ₃	0	153	-(CH ₂) ₃ -OH	1
147	-(CH ₂) ₂ -O-CH ₂ -CH ₃	0	154	-CH-(CH ₃) ₂	0
148	-(CH ₂) ₂ -S-CH ₃	0	155	-(C=O)-NH-CH ₂ -CH ₃	1
149		1	156		1
150		1			

Table 3. Substituted 4,5-Diarylisoxazoles⁸ as JNK3 Inhibitors

#	R ¹	R ²	Inhibition
166	-Br	-H	0
167	-Cl	-H	0
168	-O-CH ₂ -(CH ₃) ₂	-H	0
169	-O-(CH ₂) ₂ -C ₆ H ₅	-H	0
170	-O-CH ₂ -CH ₃	-H	0
171	-O-(CH ₂) ₂ -CH ₃	-H	0
176	-NH-(CH ₂) ₄ -CH ₃	-CH ₂ -(CH ₃) ₂	0
177	-NH-(CH ₂) ₂ -CH ₃	-CH ₂ -(CH ₃) ₂	0
178		-CH ₂ -(CH ₃) ₂	1
179		-CH ₂ -(CH ₃) ₂	0

samples with absorption, distribution, metabolism, and excretion molecular properties using QikProp.²⁵ These predicted properties are also listed in the SD files.

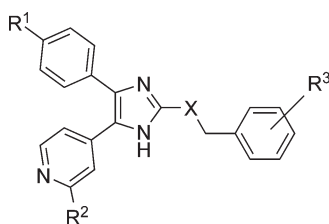
As a consequence of using only topological information in 2D-QSAR, we removed duplicates, resulting from *R*- or *S*-stereoisomers. The data set for the 2D-QSAR calculation contains 313 different compounds after preparation; 243 of these are inactive and 70 are active JNK3 inhibitors.

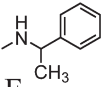
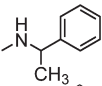
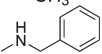
In the 3D data set, all stereoisomers without measured activities were deleted. The data set for the 3D-QSAR contains, after preparation and cleaning, 249 compounds including 189 inactive and 60 actives.

Due to the absence of IC₅₀ values, the compounds were divided in an active and an inactive class. Active compounds show a residual activity of ≤20% of the control activity. The inactive compounds have a residual activity <20%. The activity was measured in a homogeneous assay, in which the compound concentration was 10 μM, and the DMSO concentration amounted to 1%. The assay was conducted by ProQinase GmbH [ProQinase GmbH, Breisacher Str. 117, D-79106 Freiburg, Germany; <http://www.proqinase.com/>].

Structural Diversity Analysis. Structural and chemical diversity plays an important role in the field of machine learning of quantitative SAR as well as molecular modeling. For example, the information of a diversity analysis can be used to assess the generalization performance of different machine learning algorithms in leave-cluster-out validations.

To quantify the diversity of the data set, we performed a complete linkage clustering on the distance matrix of the structures. The distance matrix was calculated using Feature Trees 2.2 and the Match-Search algorithm.^{27,28} The Feature Trees algorithm is a similarity measure based on a reduced graph representation of the connectivity table of a structure. Consequently, the algorithm only utilizes 2D information to calculate the similarity between two structures. The nodes of the reduced graph symbolize pharmacophore patterns of the structures. The Feature Trees algorithm is comparable to other reduced graph algorithms like the Bemis and Murcko algorithm²⁹ or the Barker algorithm,³⁰ which was recently applied as automatic scaffold detection and diversity analysis tools. Figure 3 visualizes an example of the Feature Trees representation of a structure contained in the JNK3 data set.

Table 4. Benzylsulfanylimidazoles^{6,7} as JNK3 Inhibitors

#	R ¹	R ²	R ³	X	Inhibition
256	-F	-H	2-S-CH ₃	S	0
265	-Cl	-H	4-S(=O)-CH ₃	S	0
266	-Cl	-H	4-SO ₂ -CH ₃	S	0
269	-F	-Cl	-H	S	0
274	-F	-H	-H	CH=CH	1
278	-F	-H	-H	CH ₂	1
282	-F	-H	2-OH	S	0
286	-F	-H	3-OH	S	1
308	-F		-H	S	1
309	-F	-F	4-S(=O)-CH ₃	S	0
310	-F		4-S(=O)-CH ₃	S	1
311	-F		-H	S	0
312	-F	-Cl	4-S(=O)-CH ₃	S	0

We performed the complete linkage clustering, using different distance cutoffs in the range [0.1, 0.15, ..., 0.4] and added the results as property tags to the SD files. Further investigations were done using clusters, generated at a distance cutoff value of 0.25, based on the reasonable agreement between scaffolds and computed clusters. However, alternative clusters according to other distance cutoffs were added to the SD file.

Molecular Encodings for Machine Learning Experiments. To conduct the QSAR experiments, the JNK3 data sets were converted into a variety of structural fingerprints, including topological descriptors like linear fragments (paths), fragments (extended connectivity fingerprints), atom environments (Molprint2D-like), and possible pharmacophore points based fingerprints.

The atom type for the fingerprints was set to element symbol plus neighboring heavy atoms, the bond labels (if necessary) were single, double, triple, or aromatic. For the extended connectivity atom types, we used the Daylight invariant atom types.

After the generation of the features, the nominal representations were hashed to 5kbit fingerprints.

Radial Fingerprints. Radial environment fingerprints (RAD2D) describe the atom environment at a topological distance 1,2,...,l,...,d rather than the full paths containing bonds. This encoding was proposed by Bender et al.^{14,32} A shell $s(a_i)_l$ contains the canonically sorted set of topological neighbors of atom a_i atoms at a distance $t_{ij} = l$ between atoms ij . Additionally, we include the concatenation of all shells at a distance 1,2,...,l,...,d as additional features. Therefore, the resulting set of features contains $n \cdot d$ features. For the experiments, the search depth d was limited to 2.

Extended Connectivity Fingerprints. Extended connectivity fingerprints (ECFPs) are fragment-type features. The ECFP encodings are implemented as described by Rogers and Hahn.³³ The search depth was restricted to patterns of size four.

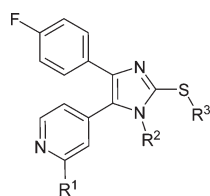
Depth-First Search Fingerprints. All-path encodings are paths, generated by a graph traversal with a modified depth-first search (DFS), as proposed by Ralaivola et al.³⁴ The linear fragments are obtained by iterating over all atoms in a molecular graph and performing an exhaustive search up to a predefined depth d . The depth d is defined by the maximum number of bonds regarded in the search. To generate a unique representation for each path, a temporary path object is generated and mapped to two nominal string representations by generating the original and reverse nominal representation of the corresponding path object. The version with the lexicographical higher order is stored. The set of features is defined as

$$F(C) = \bigcup_i^n DFS(a_i, d)$$

where $DFS(a_i, d)$ defines the local set of paths obtained by running the exhaustive depth-first search from a_i as described by Ralaivola et al. For the experiments, the search depth was limited to 8 atoms.

Pharmacophore Point Fingerprints. By this encoding, the topological or spatial relationship between a set of defined possible pharmacophore points (PPP) is described. The information of all PPPs of an atom is used to generate the fingerprint. Thus, we have to iterate over all PPPs of an atom. To keep the

Table 5. Diverse Substituted Imidazoles as JNK3 Inhibitors: Acetylaminoipyridines, [4-(Methylsulfinyl)Benzyl]Sulfanyl-Substituted Imidazole Compounds, Benzylsulfanyl Imidazoles, Methylsulfinyl Imidazoles, and Morpholine-Substituted Imidazoles^{6–8,10,31}



#	R ¹	R ²	R ³	Inhibition
91	-H		-CH ₃	0
124	-N-CH ₂ -CH ₃	-CH ₃	-CH ₃	1
126		-CH ₃	-CH ₃	1
127		-CH ₃	-CH ₃	0
131	-N-CH ₂ -CH ₃		-CH ₃	0
188		-CH ₃		1
190		-(CH ₂) ₂ -OH	-CH ₃	1
191		-(CH ₂) ₂ -OH		0
196		-(CH ₂) ₂ -OH		0
197		-(CH ₂) ₂ -O-CH ₃		0
211			-CH ₃	1
217		-CH ₂ -CH(CH ₃)OH	-CH ₃	1
234		-(CH ₂) ₂ -O-CH ₂ -C≡CH	-CH ₃	1
238		-(CH ₂) ₂ -S-(CH ₃)	-CH ₃	1
249		-(CH ₂) ₂ -CH-(OCH ₃) ₂	-CH ₃	1
251			-CH ₃	0
254	-H	-H	-CH ₂ -C≡N	1
258	-H	-H		1
296		-H	-CH ₃	1
300		-H	-CH ₃	1

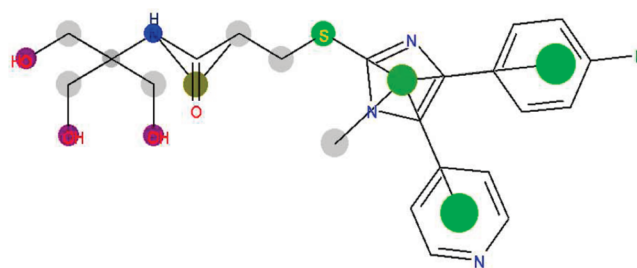


Figure 3. Feature Tree representation of the active compound 49. The circles represent nodes of the reduced graph and the color of the circles symbolize the pharmacophore type of the node.

notation simple, let P_i denote the set of valid PPPs for the i th atom. \oplus denotes a concatenation operator. Then, the set of valid pharmacophores is

$$F(C) = \bigcup_{i,j,k} P_i \oplus t_{ij} \oplus P_j \oplus t_{jk} \oplus P_k \oplus t_{ki}$$

where $t_{ij}, t_{jk}, t_{ki} \leq d$ denotes the geometrical or topological distance (either binned Euclidean distance or shortest-path distance).

The following PPP patterns were considered, as described in Renner et al.³⁵

- 1 Hydrogen-bond donor (D): [#6H] oxygen atom of an OH-group; [#7H,#7H2] nitrogen atom of an NH or NH₂ group
- 2 Hydrogen-bond acceptor (A): oxygen atom [#6]; [#7H0] nitrogen atom not adjacent to hydrogen atom
- 3 Positive (P): [*+] atom with a positive charge; [#7H2] nitrogen atom of an NH₂ group
- 4 Negative (N): [*−] atom with a negative charge; [C&\$-(C(=O)#8H1),P&\$P(=O)O,S&\$S(=O)O]] Carbon, phosphorus or sulfur atom of a COOH, POOH, or SOOH group (SMARTS replaced by a direct graph search)
- 5 Lipophilic (L): [Cl,Br,I] Chlorine, bromine, or iodine atom; [S;D2;\$S(C)(C)] Sulfur; atom adjacent to exactly two carbon atoms; carbon atom adjacent to only carbon atoms (SMARTS replaced by a direct graph search)

For the experiments, the search depth of the topological version was limited to 5, and the distance cutoff for the geometrical version was set to 6 Å. The topological version is abbreviated with PHAP2D and the geometrical version with PHAP3D. The values in Å were mathematically rounded.

Machine Learning Algorithms. In the following section, we will briefly describe several machine learning approaches that were used in the task of learning the JNK3 related activity of the compounds. The approaches can be divided into simple Bayesian approaches (Naïve Bayes), decision trees (C4.5 decision tree and random decision forest), and instance-based approaches (k nearest-neighbor classifier and support vector machines).

The source code for generating the encodings and the machine learning experiments is available. The underlying libraries for the implementation and evaluation of the machine learning tasks are open source. The fingerprint algorithms use the Chemistry Development Kit library, version 1.3.5.^{36,37} The implementations of the reference classifiers were provided by the open source machine learning library WEKA 3.71.³⁸ The experiments can be completely reproduced by the open source programs.

- Source code for the setup and validation of machine learning algorithms

- Source code for the encodings

All files necessary for the reproduction of the experiments can be downloaded at <http://www.ra.cs.uni-tuebingen.de/software/suppModJNK/>.

Naïve Bayes Classifier. The Naïve Bayesian classifier (NBC) is trained with a data set of input vectors $\mathbf{x} = (x_1, \dots, x_n)$ under the assumption that the features x_1, \dots, x_n of an input vector \mathbf{x} are independent of each other. Using Bayesian inference and the assumption that the features are independent of each other, the NBC predicts the probability $P(y_j|\mathbf{x})$ of a sample \mathbf{x} belonging to class y_j and chooses the class with highest probability

$$P(y_j|\mathbf{x}) = \frac{P(y_j) \prod_i P(x_i|y_j)}{P(\mathbf{x})}$$

Nearest-Neighbor Classifier. The k NN (k nearest neighbor classifier) computes the distance to all samples in the training set and assigns the class according to the majority class of the k nearest neighbors. Therefore, a problem of this straightforward classifier is that the computation time is completely moved to the classification step.

Decision Tree-Based Classifiers. C4.5 decision trees are iteratively constructed by evaluating the information gain of the attributes and choosing the attribute with the highest information gain as decision node for growing the tree.³⁹ This is motivated by the fact that a decision has to be based on the most informative attributes.

A random decision forest (RF) is an ensemble classifier of decision trees trained on random subspaces of the feature space. The prediction of the trained random decision forest is the consensus vote of the decision trees. As Svetnik et al.¹³ published a detailed study of random decision forests in the field of cheminformatics, we will skip further details here.

Support Vector Machines. The basic idea behind support vector machines is to separate two classes of input samples by an optimal hyperplane in a high dimensional feature space. Optimality is defined here as the maximum margin (distance) from the hyperplane to the support vectors as the border of the class distribution. With the help of a kernel function, the mapping into the high-dimensional feature space is only implicit. We evaluated two types of support vector machines in the experiments: LIBLINEAR and LIBSVM (with RBF kernel). The basic difference between these two approaches is that the complexity of LIBLINEAR is restricted to a dot product kernel but grows linear with the problem size (which enables large-scale classification tasks), whereas LIBSVM may use an arbitrary kernel but is considerably slower and has a cubic training time.

LIBLINEAR is especially suited for classifying large data sets. It was published by Fan et al.⁴⁰ in 2008. Hsieh et al.¹⁵ proposed a new coordinate descent method to solve the dual problem⁴¹ of optimizing the Lagrange multipliers, which enables to learn large problem instances in linear time. Additionally, LIBLINEAR can predict unknown instances in constant time.

LIBSVM is a nonlinear standard SVM, which was also included in this study in combination with a radial basis function kernel (RBF). The RBF kernel is suitable for the comparison of two samples $\mathbf{x}_i, \mathbf{x}_j$ with a vectorial representation. It is defined as $k_{\text{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\|\mathbf{x}_i - \mathbf{x}_j\|^2)/(2\sigma^2))$, the σ parameter gives the standard deviation of the Gaussian.

Cross-Validation. For all evaluations, the data sets were split using a seeded random number generator. Thus, all folds for all

methods are equal, making the performance of the methods comparable.

To assess the quality of the machine learning algorithms, we performed a stratified nested cross-validation for the parametrized learning algorithms. The optimal parameter combination was determined in the inner loop of the cross-validation within a 2-fold cross-validation. The results were evaluated within an outer 10-fold cross-validation. To avoid a bias induced by the initial seed for the random number generator, we conducted the validation over 10 runs (10×10 cross-validation). Therefore, each validation run resulted in 100 external evaluations on equal test sets. The results of these test sets can be compared with paired tests.

We performed a model selection to optimize the performance of the classifiers. For the support vector machines, we optimized the C parameter and the σ parameter of the radial basis function. The size of the features space for the random decision forest was also optimized, the number of trees was fixed at 100. Finally, k for k NN was optimized. Naïve Bayes is not parametrized.

We computed Matthew's correlation coefficient (MCC), the F1 measure (F1), and the area under the ROC curve (AUC) for the assessment of the models. The models were optimized for the MCC because there are more inactives than actives in the 2D and 3D benchmark data sets.

Statistical Significance. The statistical significance for the 10×10 cross-validation was evaluated by the corrected resampled t -test.⁴² The corrected resampled k -fold cross-validation test uses the following statistic

$$t = \frac{\frac{1}{k \cdot r} \sum_{i=1}^k \sum_{j=1}^r x_{ij}}{\sqrt{\left(\frac{1}{k \cdot r} + \frac{n_2}{n_1} \right) \hat{\sigma}^2}}$$

Here, for the evaluation of r runs of a k -fold cross-validation, n_1 is the number of instances used for training, n_2 is the number of instances used for testing, x_{ij} is the difference observed for fold i and run j , the variance $\hat{\sigma}^2 = (1/(k \cdot r)) \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - m)^2$, and the mean is given by $m = (1/(k \cdot r)) \sum_{i=1}^k \sum_{j=1}^r x_{ij}$. If multiple algorithms are compared, the performance is tested at a Bonferroni-corrected p -value $p_{\text{corr}} \leftarrow (p/c)$ where c is the number of classifiers in the comparison. $p \leq 0.05$ was considered to indicate a statistically significant difference.

RESULTS

Diversity Analysis. The results of the diversity analysis of the 2D and 3D data sets at the predefined distance cutoff values are compiled in Table 6. The difference between the 2D and 3D data sets stems from the reduced number of structures contained in the 3D version of the data set in comparison to the 2D version.

Furthermore, we analyzed the clusters at a distance cutoff value of 0.25. This cutoff value represents a trade-off between the number of different clusters and the average cluster size. Manual comparison with the different scaffold families, as given in the pharmaceutical background section, indicates that the different scaffolds are conserved in individual clusters. Figure 4 presents detailed information of the 12 clusters of the 2D data set at the distance cutoff value of 0.25.

Table 6. Results of the Diversity Analysis of the 2D and 3D Data Sets Using Feature Trees and a Complete Linkage Clustering^b

distance cutoff	2D		3D	
	# clusters	avg. size	# clusters	avg. size
0.15	30	10.43 ± 10.28	27 ^a	9.19 ± 7.87
0.2	19	16.47 ± 14.98	16	15.56 ± 17.80
0.25	12	26.08 ± 22.32	9	27.67 ± 20.18
0.3	6	52.17 ± 49.11	5	49.80 ± 39.88
0.35	4	78.25 ± 43.80	3	83.00 ± 22.91

^aContains one singleton cluster. ^bThe distance cutoff value represents the maximum distance value between two structures of a cluster.

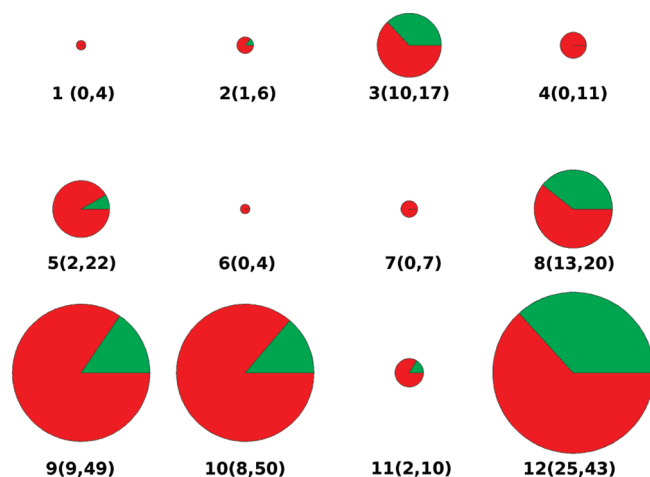


Figure 4. Visualization of the individual clusters. The size of each pie plot represents the number of structures that belongs to each cluster. The red and green fraction of the pie plots depict the distribution of active and inactive structures, respectively. The number below each pie plot represents the cluster number and the first and second value in parentheses represents the number of active and inactive structures, respectively.

The size of a cluster ranges from 4 (cluster 1 and 2) to 68 (cluster 12) structures. This diversity of the cluster size implies that the chemical space is not uniformly covered by the structures of the data set. Eight out of 12 clusters contain active structures, which indicates that the active structures are not located on one activity island, but are rather scattered among several different activity islands. Additionally, the distribution of the active structures within clusters that contain active structures ranges from 8.33% (cluster 5) to 39.39% (cluster 8).

Machine Learning. The results for the 2D benchmark using the topological encodings are presented in Tables 7–10. All results that were not significantly worse than any other reference approach are shown in bold. Therefore, there may be several “best” performing algorithms.

The random decision forest achieved the best averaged MCC performance for all encodings, ranging from 0.548 ± 0.186 , using the radial atom environments, and to 0.744 ± 0.174 using the DFS encodings. The performance of the PPP-point encoding was also convincing with 0.723 ± 0.220 . LIBLINEAR turned out to be the second-best approach ranked second using the DFS, ECFP, and the PPP encoding. *k*NN was not significantly

Table 7. Results of the Nested 10-Fold Cross-Validation for the DFS Encoding on the 2D Benchmark^a

classifier	MCC	AUC	F1
RF	0.744 ± 0.174	0.760 ± 0.214	0.434 ± 0.311
LIBLINEAR	0.686 ± 0.199	0.724 ± 0.186	0.437 ± 0.284
kNN	0.657 ± 0.209	0.707 ± 0.194	0.437 ± 0.268
NBC	0.622 ± 0.222	0.681 ± 0.202	0.454 ± 0.260
C4.5	0.609 ± 0.243	0.649 ± 0.248	0.470 ± 0.286
LIBSVM	0.351 ± 0.146	0.727 ± 0.213	0.435 ± 0.306

^aThe models for the parameterized methods were optimized according to the MCC. Significantly best performing methods are indicated by bold font.

Table 8. Results of the Nested 10-Fold Cross-Validation for the ECFP Encoding on the 2D Benchmark^a

classifier	MCC	AUC	F1
RF	0.693 ± 0.186	0.750 ± 0.187	0.434 ± 0.283
LIBLINEAR	0.685 ± 0.214	0.736 ± 0.159	0.406 ± 0.269
kNN	0.620 ± 0.213	0.690 ± 0.178	0.449 ± 0.241
C4.5	0.620 ± 0.254	0.690 ± 0.160	0.397 ± 0.227
NBC	0.604 ± 0.245	0.729 ± 0.130	0.414 ± 0.217
LIBSVM	0.359 ± 0.125	0.727 ± 0.197	0.423 ± 0.294

^aThe models for the parameterized methods were optimized according to the MCC. Significantly best performing methods are indicated by bold font.

Table 9. Results of the Nested 10-Fold Cross-Validation for the PHAP2D Encoding on the 2D Benchmark^a

classifier	MCC	AUC	F1
RF	0.723 ± 0.220	0.764 ± 0.197	0.408 ± 0.294
LIBLINEAR	0.702 ± 0.199	0.744 ± 0.168	0.419 ± 0.283
C4.5	0.663 ± 0.205	0.707 ± 0.174	0.431 ± 0.258
kNN	0.647 ± 0.239	0.721 ± 0.174	0.410 ± 0.249
NBC	0.607 ± 0.215	0.685 ± 0.178	0.449 ± 0.236
LIBSVM	0.368 ± 0.132	0.734 ± 0.219	0.420 ± 0.311

^aThe models for the parameterized methods were optimized according to the MCC. Significantly best performing methods are indicated by bold font.

Table 10. Results of the Nested 10-Fold Cross-Validation for the RAD2D Encoding on the 2D Benchmark^a

classifier	MCC	AUC	F1
RF	0.548 ± 0.186	0.829 ± 0.094	0.759 ± 0.097
LIBSVM	0.545 ± 0.195	0.736 ± 0.101	0.750 ± 0.106
LIBLINEAR	0.497 ± 0.201	0.742 ± 0.103	0.741 ± 0.101
kNN	0.449 ± 0.192	0.766 ± 0.127	0.717 ± 0.096
C4.5	0.428 ± 0.186	0.704 ± 0.092	0.707 ± 0.093
NBC	0.325 ± 0.223	0.757 ± 0.119	0.655 ± 0.111

^aThe models for the parameterized methods were optimized according to the MCC. Significantly best performing methods are indicated by bold font.

worse on the ECFP and PPP encoding. The remaining approaches were outperformed significantly on at least three of four benchmarks.

Table 11. Results of the Nested 10-Fold Cross-Validation for the PHAP3D Encoding on the 3D Benchmark^a

classifier	MCC	AUC	F1
LIBSVM	0.524 ± 0.231	0.756 ± 0.116	0.750 ± 0.120
LIBLINEAR	0.517 ± 0.219	0.762 ± 0.111	0.746 ± 0.114
NBC	0.495 ± 0.206	0.760 ± 0.122	0.733 ± 0.107
C4.5	0.488 ± 0.205	0.751 ± 0.115	0.734 ± 0.103
RF	0.469 ± 0.211	0.805 ± 0.110	0.724 ± 0.109
kNN	0.437 ± 0.241	0.754 ± 0.144	0.706 ± 0.123

^aThe models for the parameterized methods were optimized according to the MCC. Significantly best performing methods are indicated by bold font.

The results for the geometrical three-point PPP encoding are presented in Table 11. This benchmark was solved best by support vector machines with an average MCC value in between 0.517 ± 0.219 and 0.524 ± 0.231 . The NBC, C4.5, and RF were not significantly worse.

The overall performance considering all machine learning approaches on the 3D benchmark was found to be worse compared to the 2D version. This indicates that this benchmark is a more difficult learning task.

DISCUSSION AND CONCLUSION

In this study, we presented and described a novel 2D- and 3D-QSAR data set for JNK3 inhibitors as a binary classification problem. The compounds were assigned to a 2D and 3D version. To provide an unambiguous data set, we removed the stereoisomers from the 3D data set, if there are no activities available for both isomers. To analyze the diversity, we conducted clustering experiments with Feature Trees to determine the number of clusters at different distance cutoffs. Additionally, we investigated whether this benchmark can be learned with straightforward 2D- and 3D-QSAR approaches.

The diversity of this data set was quantified by a complete linkage clustering on the distance matrix of the structures, using Feature Trees. The most appropriate clusters were found at a distance cutoff of 0.25. These clusters were regarded in a further analysis. The diversity analysis showed that the clusters do not uniformly cover the chemical space. The different distributions of the active compounds represent a challenging task for machine learning and molecular modeling algorithms.

With a variety of molecular encodings and machine learning approaches, we evaluated the modeling performance on the 2D and 3D version of the data set. We applied machine learning approaches like random decision forest, Naïve Bayes, *k* nearest neighbor, linear and nonlinear support vector machines, and C4.5 decision trees. As molecular encodings, we employed radial fingerprints, extended connectivity fingerprints, depth-first search fingerprints, and pharmacophore point fingerprints. With these approaches, we were able to show that the described problem can be learned with convincing results. The setup for the evaluation of the machine learning algorithms was designed to fully avoid overfitting because the parameters were optimized in a nested 10-fold cross-validation. The random decision forest achieved on the 2D benchmark the best results with an averaged MCC of 0.744 ± 0.174 using the DFS encoding. The MCC values on the 3D benchmark reached up to an average MCC of 0.524, using a support vector machine and a pharmacophore fingerprint, which

indicates that this is a more difficult learning task. As the results were obtained, using default encodings without a further tuning of parameters, such as search depth or distance cutoffs, it should be possible to improve these results further.

It might be considered as a drawback that we just present a binary classification problem, which stems from the absence of IC₅₀ values. However, the data set can be employed to calibrate alternative molecular modeling approaches like scoring functions for molecular docking, pharmacophore point extraction, virtual screening approaches, or to investigate basic properties of the compounds such as lipophilicity or charges. It becomes clear from the machine learning results and the diversity analysis that the actives cannot be trivially separated from the inactives by obvious structural similarities.

To conclude, we provide a binary 2D- and 3D-QSAR benchmark for future research comprising a diverse set of JNK3 inhibitors. Based on the results of our cluster analysis and QSAR models, we think that these data are an interesting contribution to the research field. To the best of our knowledge, our data set is the first public data set for JNK3 inhibitors with measured activities from the same assay.

ASSOCIATED CONTENT

S Supporting Information. The 2D and 3D version for the JNK3 inhibitors plus their measured activities. The files are annotated with descriptors provided by QikProp from Schrödinger and the clusters computed by Feature Trees. The 2D version and 3D version of the described data set annotated with feature tree clusters are provided. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49(0)7071-2972459. Fax: +49 (0)7071-295037.
E-mail: stefan.laufer@uni-tuebingen.de.

REFERENCES

- (1) Kyriakis, J.; Avruch, J. Mammalian Mitogen-Activated Protein Kinase Signal Transduction Pathways Activated by Stress and Inflammation. *Physiol. Rev.* **2001**, *81*, 807.
- (2) Siddiqui, M.; Reddy, P. Small Molecule JNK (c-Jun N-Terminal Kinase) Inhibitors. *J. Med. Chem.* **2010**, *53*, 3005–3012.
- (3) Shaikh, A.; Ismael, M.; Del Carpio, C.; Tsuboi, H.; Koyama, M.; Endou, A.; Kubo, M.; Broclawik, E.; Miyamoto, A. Three-dimensional quantitative structure-activity relationship (3D-QSAR) and docking studies on (benzothiazole-2-yl) acetonitrile derivatives as c-Jun N-terminal kinase-3 (JNK3) inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, S917–S925.
- (4) Ijjaali, I.; Petitet, F.; Dubus, E.; Barberan, O.; Michel, A. Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. *Bioorg. Med. Chem.* **2007**, *15*, 4256–4264.
- (5) Chung, J.; Cho, A.; Hah, J. Hologram and Receptor-Guided 3D QSAR Analysis of Anilinoipyridine JNK3 Inhibitors. *Bull. Korean Chem. Soc.* **2009**, *30*, 2739–2748.
- (6) Laufer, S.; Wagner, G.; Kotschenreuther, D.; Albrecht, W. Novel Substituted Pyridinyl Imidazoles as Potent Anticytokine Agents with Low Activity against Hepatic Cytochrome P450 Enzymes. *J. Med. Chem.* **2003**, *46*, 3230–3244.

- (7) Laufer, S.; Striegel, H.; Wagner, G. Imidazole Inhibitors of Cytokine Release: Probing Substituents in the 2 Position. *J. Med. Chem.* **2002**, *45*, 4695–4705.
- (8) Laufer, S.; Margutti, S.; Fritz, M. Substituted Isoxazoles as Potent Inhibitors of p38 MAP Kinase. *ChemMedChem* **2006**, *1*, 197–207.
- (9) Laufer, S.; Margutti, S. Isoxazolone Based Inhibitors of p38 MAP Kinases. *J. Med. Chem.* **2008**, *51*, 2580–2584.
- (10) Laufer, S.; Zimmermann, W.; Ruff, K. Tetrasubstituted Imidazole Inhibitors of Cytokine Release: Probing Substituents in the N-1 Position. *J. Med. Chem.* **2004**, *47*, 6311–6325.
- (11) Laufer, S.; Domeyer, D.; Scior, T.; Albrecht, W.; Hauser, D. Synthesis and Biological Testing of Purine Derivatives as Potential ATP-Competitive Kinase Inhibitors. *J. Med. Chem.* **2005**, *48*, 710–722.
- (12) *Schrödinger LigPrep, version 2.2*; Schrödinger, LLC: New York, NY, 2005.
- (13) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
- (14) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
- (15) Hsieh, J.-H.; Wang, X. S.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification k NN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 593–609.
- (16) Boldt, S.; Kolch, W. Targeting MAPK Signalling: Prometheus' Fire or Pandora's Box?. *Curr. Pharm. Des.* **2004**, *10*, 1885–1905.
- (17) Manning, G.; Whyte, D.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912.
- (18) Traxler, P.; Furet, P. Strategies toward the Design of Novel and Selective Protein Tyrosine Kinase Inhibitors. *Pharmacol. Ther.* **1999**, *82*, 195–206.
- (19) Scapin, G.; Patel, S.; Lisnock, J.; Becker, J.; LoGrasso, P. The Structure of JNK3 in Complex with Small Molecule Inhibitors: Structural Basis for Potency and Selectivity. *Chem. Biol. (Cambridge, MA, U. S.)* **2003**, *10*, 705–712.
- (20) *Schrödinger Suite 2010, Induced Fit Docking protocol*; Schrödinger, LLC: New York, NY, 2010.
- (21) Vesely, J.; Havlicek, L.; Strnad, I. Inhibition of Cyclin-Dependent Kinases by Purine Analogues. *Eur. J. Biochem.* **1994**, *224*, 771–786.
- (22) Wang, Z.; Canagarajah, B.; Boehm, J.; Kassisa, S.; Cobb, M.; Young, P.; Abdel-Meguid, S.; Adams, J.; Goldsmith, E. Structural basis of inhibitor selectivity in MAP kinases. *Structure (Cambridge, MA, U. S.)* **1998**, *6*, 1117–1128.
- (23) Lisnock, J.; Griffin, P.; Calaycay, J.; Frantz, B.; Parsons, J.; O'Keefe, S.; LoGrasso, P. Activation of JNK3 [alpha] 1 Requires both MKK4 and MKK7: Kinetic Characterization of in Vitro Phosphorylated JNK3 [alpha] 1. *Biochemistry* **2000**, *39*, 3141–3148.
- (24) *Schrödinger Maestro, version 8.5*; Schrödinger, LLC: New York, NY, 2008.
- (25) *Schrödinger QikProp, version 3.3*; Schrödinger, LLC: New York, NY, 2010.
- (26) Hauser, D.; Scior, T.; Domeyer, D.; Kammerer, B.; Laufer, S. Synthesis, Biological Testing, and Binding Mode Prediction of 6, 9-Diaryl-purin-8-ones as p38 MAP Kinase Inhibitors. *J. Med. Chem.* **2007**, *50*, 2060–2066.
- (27) Rarey, M.; Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
- (28) Rarey, M.; Hindle, S.; Maaß, P.; Metz, G.; Rummey, C.; Zimmermann, M. In *Pharmacophores and Pharmacophore Searches, Pharmacophores and Pharmacophore Searches*; Langer, T., Hoffmann, R., Eds.; Wiley-VCH: Weinheim, Germany, 2006; Chapter Feature trees: Theory and application from large-scale virtual screening to data analysis, pp 81–116.
- (29) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (30) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346–356.
- (31) Laufer, S.; Wagner, G. From Imidazoles to Pyrimidines: New Inhibitors of Cytokine Release. *J. Med. Chem.* **2002**, *45*, 2733–2740.
- (32) Bender, A.; Mussa, H. Y.; Glen, R. C. Screening for Dihydrofolate Reductase Inhibitors Using MOLPRINT 2D, a Fast Fragment-Based Method Employing the Naive Bayesian Classifier: Limitations of the Descriptor and the Importance of Balanced Chemistry in Training and Test Sets. *J. Biomol. Screen.* **2005**, *10*, 658–666.
- (33) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.
- (34) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093–1110.
- (35) Renner, S.; Fechner, U.; Schneider, G. In *Pharmacophores and Pharmacophore Searches, Pharmacophores and Pharmacophore Searches*; Langer, T., Hoffmann, R., Eds.; Wiley-VCH: Weinheim, Germany, 2006; Chapter Alignment-free Pharmacophore Patterns - A Correlation Vector Approach, pp 49–79.
- (36) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- (37) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.
- (38) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18.
- (39) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Francisco, 1993.
- (40) Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; Lin, C.-J.; LIBLINEAR, A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
- (41) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, USA, 2002.
- (42) Bouckaert, R. R.; Frank, E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining - Proceedings of 8th Pacific-Asia Conference, PAKDD 2004*; Dai, H., Srikant, R., Zhang, C., Eds.; Springer: Heidelberg, 2004; Vol. 3056, pp 3–12.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published on the Web on January 31, 2011, with an error. In the Materials and Methods section, the version listed for the Chemistry Development Kit was incorrect. The corrected version was reposted on February 21, 2011.