JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# Exploiting Structural Information in Patent Specifications for Key Compound Prediction

Christian Tyrchan,*[,†] Jonas Boström,[†] Fabrizio Giordanetto,[†] Jon Winter,[§] and Sorel Muresan[‡]

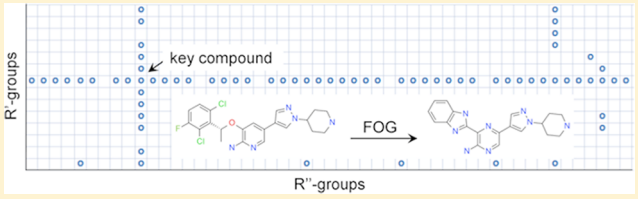
[†]AstraZeneca R&D, CVGI iMed, Pepparedsleden 1, S-431 83 Mölndal, Sweden

[‡]AstraZeneca R&D, Discovery Sciences Computational Sciences, Pepparedsleden 1, S-431 83 Mölndal, Sweden

[§]AstraZeneca R&D, Oncology iMed, Alderley Park, Macclesfield, Cheshire, SK10 4TG, United Kingdom

Supporting Information

**ABSTRACT:** Patent specifications are one of many information sources needed to progress drug discovery projects. Understanding compound prior art and novelty checking, validation of biological assays, and identification of new starting points for chemical explorations are a few areas where patent analysis is an important component. Cheminformatics methods can be used to facilitate the identification of so-called key compounds in patent specifications. Such methods, relying on structural information extracted from documents by expert curation or text mining, can complement or in some cases replace the traditional manual approach of searching for clues in the text. This paper describes and compares three different methods for the automatic prediction of key compounds in patent specifications using structural information alone. For this data set, the cluster seed analysis described by Hattori et al. (Hattori, K.; Wakabayashi, H.; Tamaki, K. Predicting key example compounds in competitors' patent applications using structural information alone. *J. Chem. Inf. Model.* **2008**, *48*, 135−142) is superior in terms of prediction accuracy with 26 out of 48 drugs (54%) correctly predicted from their corresponding patents. Nevertheless, the two new methods, based on frequency of R-groups (FOG) and maximum common substructure (MCS) similarity measures, show significant advantages due to their inherent ability to visualize relevant structural features. The results of the FOG method can be enhanced by manual selection of the scaffolds used in the analysis. Finally, a successful example of applying FOG analysis for designing potent ATP-competitive AXL kinase inhibitors with improved properties is described.




## INTRODUCTION

The chemical information landscape is changing rapidly with a yearly increase of over 1 million new compounds and more than 700,000 publications related to chemistry.[2−4] Accordingly, the chemical space claimed by patents is becoming more crowded and competitive, especially for target classes such as protein kinases and G-protein-coupled receptors (GPCRs).[5] Understanding the chemical space covered by relevant journals and patent specifications is an important activity in medicinal chemistry projects to support business decisions.

Patent informatics or "patinformatics" combines text mining with computational analysis to identify relevant information from published documents. The process has been defined as macroscopic (between documents) and microscopic (inside a document) depending on the scale of analysis.[6,7] The information extracted from patent specifications (the term "patent specification" is used throughout the manuscript as a general reference to a patent text be it an application or a granted patent), including chemistry, is usually registered in databases from which it can be subsequently accessed, further processed, and visualized.[8−11] Such tools have been primarily used by information scientists, and large scale analyses or method development in the field of patent informatics is still a growing area.[12−14]

In patent specifications, the claimed chemical space is usually defined by generic Markush structures, which are exemplified by a number of synthesized compounds and their associated biological profiles. Extracting chemical entities from these documents is not a trivial task and different approaches are currently used. These approaches include manual extraction by expert curators, automatic text mining techniques supported by chemical named entity recognition (CNER), or combinations thereof.[15] A variety of commercial institutions now supply databases with high quality, manually extracted structures of exemplified compounds.[16−18] Following the advances in text mining technologies and improvements of the CNER process, several databases containing automatically extracted chemical entities from patent specifications are also available.[19,20]

At AstraZeneca, exemplified compounds and corresponding annotations, intermediates, and reagents, as well as prophetic compounds (potential chemical variations not yet synthesized) are stored in relational databases allowing for complex chemical and text queries.[21] Consequently, computational methods can be applied for data analysis and visualization to support medicinal chemistry teams in evaluating the available information. In particular, the identification of key com-

**Received:** March 12, 2012
**Published:** May 29, 2012

pound(s) in a patent specification in order to facilitate experimentation with the subject matter of that invention is important—they may help to validate experimental protocols, as well as assist in understanding the limits of the physicochemical, pharmacokinetic, and pharmacodynamic (PK/PD) properties and clinical profiles. Furthermore, identification of key compounds from a patent specification may provide a starting point for novel lead finding activities. Associated biological activity data can be useful for medicinal chemistry projects on similar targets or to improve secondary pharmacology predictions. The current practice of identifying key compounds, a manual process where one searches for clues in the text (e.g., scale of reaction, powder diffraction data, additional biological in vitro and in vivo data), is time-consuming and tedious. Cheminformatics tools and access to large patent databases now offer opportunities to automate and facilitate the process.[22]

**What Is a Key Compound?** When faced with a medicinal chemistry patent specification, one of the first questions a researcher typically asks is "What is the key compound in this document?" Depending on the nature of the patent or patent application, the nature of the "best" compound will vary. If the document is a patent that claims a series of compounds which have entered clinical development, the "best" compound will most likely be the drug candidate. If the document is a patent application covering compounds at an earlier stage of the discovery process, the "best" compound may be the compound with the optimal physicochemical properties, the most biologically active tool or probe, or the most suitable pharmacokinetic profile for the desired indication. In either of these cases, it is useful for a researcher to be able to identify a small number of compounds, and preferably a single compound. This can then be experimented with more closely.

A patent application is an agreement between inventor and state, allowing an inventor a monopoly over their invention for a limited time in return for disclosure of their novel invention. In the EPC (European Patent Convention), applicants are required to disclose their inventions in a manner sufficiently clear and complete for them to be carried out by a person skilled in the art.[23] In the United States, inventors are additionally required to include the best mode of making or practicing the invention. The implication of this is that an inventor of a medicinal chemistry patent application must include the "best" compound(s), e.g., that or those which show particularly favorable properties, including clinical candidates. Failure to include the "best" compounds, if they are thought to be the best available at the time of filing of the application, has been a potential ground for invalidation of a patent in the United States.[24] While the best mode compound must be included in the patent specification, there is, however, no requirement for the inventor to draw attention to its identity. The responsibility therefore falls on the reader of the document to infer the best mode compound by careful study of the text for clues. Frequently, such key compounds have been synthesized on a larger scale than other examples, are described in multiple salt forms or formulations, have been tested in more complex biological assays, are presented with animal pharmacokinetic data, or are specifically named in the claims section of the document.

Searching for the clues that identify a key compound is a time-consuming and tedious process because patent specifications can run over many hundreds of pages and are not necessarily published in a language the reader may understand. Compound examples may be described using complex IUPAC names, which require interpretation, rather than depicted as chemical structures. While knowledge engineering and text mining approaches to identify key compounds in a similar manner to the manual methods outlined above may one day be available, in their absence a different automated approach is desirable. Furthermore, clues may not even be present. Overall, a method that enables the identification of key compounds in the absence of supporting metadata is useful for the assessment of patent specifications. In the present study, three different methods for key compound prediction, two of them novel, based on structural information alone are compared: the cluster seed approach (CSA) described by Hattori et al.,[1] the Molecular Idol, a derivative approach using MCS-based similarities, and the frequency of R-group analysis (FOG). In addition, an example of applying FOG analysis in a medicinal chemistry project is presented.

**Data Set.** The data set of patent specifications used to assess the ability of the three methods to predict key compounds was generated by the following procedure.

As a starting point, we used the data set published by Hattori et al. consisting of 30 drugs and their earliest associated patent.[1] They were originally selected from a list of U.S. 2005 top-selling drugs with the premise that the corresponding patent contains more than 50 example compounds. This data was extended with a set of 18 drugs from the GVKBIO drug database[25] for which the corresponding earliest patents were retrieved from Thomson Reuters Pharma.[17] All structures related to a patent were retrieved from the GVKBIO Target Class databases.[18] It should be noted that GVKBIO uses expert curators to populate the Target Class databases with exemplified compounds following clearly defined business rules.

Overall, a total of 48 drugs were identified with their associated first patent specifications. The number of compounds in these patents ranged from 10 to 510, with a median of 63 and a mean of 94 compounds (see Table 1 in the Supporting Information). All structures were standardized with Pipeline Pilot 7.5 as follows: adducts (counterion or water) were removed, charges were neutralized, stereochemistry was removed, and the canonical tautomer was generated.[26]

**Key Compound Prediction Methods.** The cluster seed analysis (CSA) published by Hattori et al. was reimplemented in order to compare it with the two new approaches for key compound prediction that have been developed: Molecular Idol and Frequency of Group (FOG). All three methods use structural information alone and are based on the assumption that medicinal chemists carry out extensive chemistry around the key compound (i.e., the marketed drug in this study). The main difference between the three methods is the way they define series or clusters from a set of compounds.

**Cluster Seed Analysis.** In the CSA approach, a fingerprint-based nearest neighbor analysis is used to identify cluster centers in the patent chemical space, and key compounds are assumed to be located at the centers of densely populated regions (clusters). Compounds are ranked on the basis of the number of their neighbors, with key compounds displaying the highest number of neighbors. The method was reimplemented in PipelinePilot 7.5 using ECFP4 fingerprints with a 0.7 Tanimoto cutoff, in line with the original work.[1]

**Molecular Idol.** Similar to CSA, Molecular Idol aims at identifying compounds that have a large number of near neighbors within a given set of compounds. In contrast to CSA, Molecular Idol is based on a maximum common substructure

(MCS) similarity measure instead of a fingerprint similarity measure.[27] In general, fingerprint-based similarity measures consider properties of entire molecules rather than distinct scaffolds. This is a potential drawback because the distinct scaffolds are often seen as the basis of medicinal chemistry projects, for example, when exploring structural series. Additionally, such a distinct scaffold is more than a common subgraph shared by a family of molecules. It often embodies a specific synthetic strategy that allows systematic exploration of the structure activity relationship (SAR) space. Although MCS perception is widely used in establishing SAR from 2D chemical structures, its use in large scale searching and clustering has been limited by the performance of the MCS algorithms. Recent developments have alleviated this obstacle.[28,29]

When running Molecular Idol, one has the option to define a similarity threshold value (TanimotoMCSS) determining if two compounds are considered to be near neighbors or not. In this analysis, the TanimotoMCSS was set to the default value of 0.9. Subsequently, an all-against-all comparison is performed, and the count of near neighbors exceeding the given threshold is recorded for each compound in the set. Such near neighbors are termed fans, and the compounds are ranked according to their number of fans. The compound with the highest number of fans is predicted to be the Molecular Idol or, in other words, the key compound. In the event of a tie (the same number of fans), the compounds are ranked by their accumulated TanimotoMCSS score (the sum of the TanimotoMCSS scores for the defined near neighbors).

Molecular Idol was implemented using the OEChem toolkit and Python. Each molecule in the given data set was transformed into a molecular graph (OEGraphMols), assigning aromaticity and suppressing hydrogens. The OEMCSSearch function was used for the maximum common substructure searches in the default mode, which match atoms having the same atomic number, aromaticity, and formal charge as matching bonds with the same order (taking aromaticity into account). The time-effective "approximate" search option was used in all cases. The parameters that could potentially affect the results were varied in an exhaustive fashion to increase the probability of accurately predicting the key compound within the top 5 ranked compounds. The different experiments showed that our method performed consistently well, provided that the similarity measure (TanimotoMCSS) was set to a high value (>0.8) (results not shown).

**Frequency of Group Analysis.** Fragment occurrence analysis has been used to transfer knowledge between chemical series, for example, in the context of fragments or scaffold hopping.[30] The FOG approach is based on the assumption that chemists iteratively identify favorable side chains (i.e., R-groups) in a chemical series and use them more often in subsequently designed sets. Here, the frequency of occurrence of a given R-group is taken as an indication of how favorable a side chain is for the overall compound profile. Accordingly, key compounds should be found at the intersections of frequently occurring R-groups.

A new algorithm for automatic series detection and R-group analysis was developed using PipelinePilot 7.5 (Figure 1). Molecular frameworks are generated (Murcko Assemblies, excluding alpha atoms, default settings in the PipelinePilot corresponding protocol) and clustered (ECFP6 fingerprints, maximum distance 0.6, nearest outside maximum) to enable a more optimal pregrouping of possible series.[31] The largest maximal subgraph for each cluster is identified and ranked by

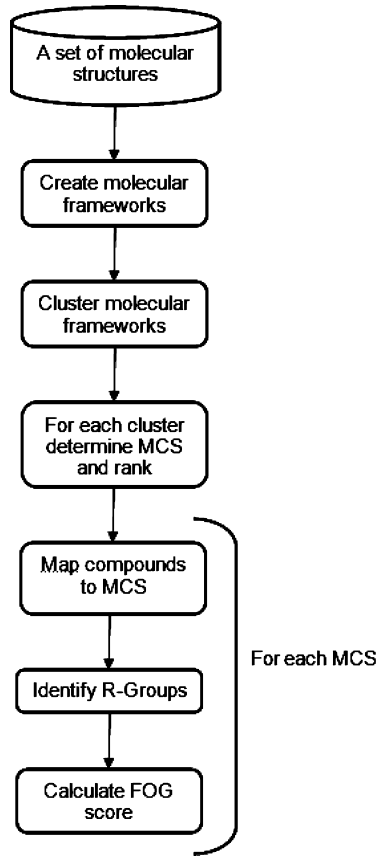



**Figure 1.** Schematic display of the FOG algorithm with automatic series extraction.

significance. This is determined by calculating all pairwise Tanimoto similarities between the largest maximal subgraphs using ECFP4 fingerprints with no Tanimoto cutoff. The scaffolds are ranked accordingly from the highest to the lowest number of neighbors. Scaffolds with no neighbors or the same number of neighbors are ranked by the number of heavy atoms (high to low) to favor larger scaffolds over smaller ones. The scaffolds are then mapped back onto the compound set in the order of their ranking, ensuring that each compound is assigned to only one scaffold. An R-group analysis is performed, and the most frequently occurring R-groups in the scaffold series are identified. A FOG-score is then calculated for each compound assigned to a scaffold by adding the corresponding occurrence of each R-group. As default setting, the two highest scoring compounds for each of the first five scaffolds are reported back.

## RESULTS AND DISCUSSION

Medicinal chemistry projects typically select between one to three key compounds for experimentation, e.g., advanced testing and profiling because of the nature of the tests that are subsequently performed. In general, a project team attempts to retrieve or generate as much information as possible to differentiate and optimize new compounds. Such information usually includes in vitro or in vivo activity, physicochemical properties (e.g., lipophilicity, solubility), PK/PD, ADME, and toxicity profiles. Many of these tests are costly and time consuming. Therefore, five selected key compounds seem a reasonable initial set to be used for a final manual inspection.

All three methods used in the current work rank compounds in terms of number of neighbors or number of scaffolds. The

prediction accuracy alone, determined by the number of times a key compound was correctly predicted within a given number of ranked compounds, is a rather imprecise measure because more than one compound can be assigned to the same rank. In the worst case scenario, all compounds from a patent specification can be equally ranked. Therefore, two additional quality criteria are introduced to address this issue: the number of key compounds correctly predicted on the first rank and the number of times the predicted key compound is among the predicted top 5 compounds in contrast to the top 5 ranks. The performance of the three methods is summarized in Table 1. As

**Table 1. Results of the CSA, Molecular Idol, and FOG Methods on the 48 Drugs Set**

| Method | No. of times predicted as the first ranked compound | No. of times predicted within the first five ranked compounds | No. of times compounds to drug $\leq 5^a$ |
|---|---|---|---|
| CSA | 11 (23%) | 26 (54%) | 73% (19/26) |
| Molecular Idol | 5 (10%) | 22 (46%) | 100% (22/22) |
| FOG | 11 (23%) | 17 (35%) | 88% (15/17) |

[a]Number of times the real key compound is within the first five predicted key compounds in contrast to the first five ranks.

listed in Table 1, the number of key compounds correctly predicted on the first rank as quality criteria is indeed not enough to judge the quality of the prediction. In the case of CSA, in 26 cases, the key compound was found within the first five ranks, but in only 19 of these cases only one compound was actually assigned to a rank. In the other cases, more than five compounds where assigned to the five ranks (Table 2).

**Comparison of the Data Sources.** The selected data set consists of 48 patents with associated drugs having more than 10 compounds in the GVKBIO database, containing the drug set used in the analysis published by Hattori et al.[1] In each of these 30 cases, the corresponding patent specification contains the key compound, the actual drug on the market, but with a significantly different number of exemplified compounds in the chosen patent specification compared to the number of extracted compounds from each of the databases. This can be explained on one hand by the different business rules for extraction of compounds from patents followed by the two providers. GVKBIO extracts only exemplified compounds (including reference compounds) with distinct biological activities from English language patent specifications. Chemical Abstracts Service (CAS) includes all "relevant" compounds including exemplified compounds, reagents, intermediates, and prophetic compounds for any patent specification.

Additionally, the used patent specification in the analysis may differ. In only 13 out of 30 cases, GVKBIO has abstracted the same patent specification as used by Hattori et al. in their analysis.[1]

Despite the availability of the drugs Atacand and Patanol in the GVKBIO database, these were the only two cases from the Hattori set that were not included in our analysis.

The GVKBIO database contains four compounds assigned to the granted Atacand European Patent EP0459136(B1) (published in 2002 with 8 exemplified compounds), whereas Hattori et al. reported 67 examples for the published European Patent Application EP0459136(A1) (published in 1996 with 163 working examples). For Patanol, seven compounds from the International Patent Application WO1996039147 (publication date 1996) are retrieved from the GVKBIO database. From these, two are actually reference compounds (Cromolyn sodium and Nedocromil sodium), and Patanol is exemplified as *cis/trans* isomer. Hattori et al. reported 151 compounds from European Patent Application EP235796(A2) (publication date 1997), which contains 64 exemplified compounds. Furthermore, the European Patent Application EP235796(A2) is not derived from the International Patent Application WO1996039147.

The consequences of the differences found in the two data sources are two-fold. First, the real key compound (the possible drug or drug candidate) does not necessarily have to be contained in either of the two data sources, and the prediction methods will all fail. Second, it cannot be expected that the different data sources contain all exemplified compounds from the patent specification or are even from the same patent specification or patent family. Therefore, the chemical space covered by the patent specification is not necessarily equally represented by the two data sources. This could be partially reflected by the comparison of the results from our in-house cluster seed method and the published results,[1] as discussed below.

**Performance of the Cluster Seed Method.** Hattori et al. reported a 57% (17 out of 30) prediction accuracy for identifying the drug within the first five predicted key compounds. Using our implementation of the cluster seed method, the drug is selected in 26 cases (prediction accuracy 54%), of which 18 (64%) are from the Hattori set. From these 18 drugs, five are not correctly predicted by Hattori et al. On the other hand, using our implementation, three drugs were not identified as key compounds (Table 2).

The reasons for these differences in the prediction of key compounds between Hattori et al. and our work are not obvious. One explanation could be that despite a careful implementation of every detail of the published algorithm subtle differences may still exist. Another source of discrepancy is the data set. As discussed above, the number of compounds per patent specification, the used patent specification as such, and consequently the composition of the compound sets from the different sources can differ significantly (Table 1 in the Supporting Information). Incidentally, no single patent specification has the same number of compounds from the two data sources, and in 27 cases, respectively in two out of 13 cases where the same patent specification is found, the GVKBIO database contains more compounds (i.e., more exemplified compounds) compared to the data published by Hattori et al.

**Performance of Molecular Idol.** Molecular Idol correctly predicts 22 key compounds out of 48 cases. Figure 2 depicts the application interface and result page. It is shown that Raxar, an oral broad-spectrum quinoline antibacterial agent, is predicted among the first five key compounds. The corresponding first patent specification, WO89006649, is a typical example of SAR exploitation. GVKBIO extracted 100 structures from this document, and as many as 26 compounds show a TanimotoMCSS greater than 0.9 when compared to the key compound Raxar. This means that those structures cannot differ from Raxar by more than 2 atoms and/or that the MCS cannot include fewer than 24 of Raxar's 26 heavy atoms. Neither CSA nor FOG predicted Raxar as a key compound.

In contrast to Raxar, Avapro, an angiotensin II antagonist mainly used for the treatment of hypertension, was not predicted to be a key compound by Molecular Idol. The reason

**Table 2. Detailed Key Compound Prediction Results for the 35 Drugs Correctly Predicted by at Least One of the Three Methods**

| Drug | Patent Specification | CSA[a] | FOG[a] | Molecular Idol[a] | Prediction Hattori et al.[a] | No. of compounds to drug CSA[b] | No. of compounds to drug FOG[b] | No. of compounds to drug Molecular Idol[b] |
|---|---|---|---|---|---|---|---|---|
| Aciphex | EP0268956 | 1 | 1 | 1 | 1 | 2 | 0 | 2 |
| Aldara | EP145340 | 1 | 0 | 0 | 0 | 9 | | |
| Aricept | EP0296560 | 1 | 1 | 1 | 1 | 1 | 6 | 1 |
| Arimidex | EP0296749 | 1 | 1 | 1 | 1 | 1 | 0 | 2 |
| Avapro | WO1991014679 | 0 | 1 | 0 | 0 | | 2 | |
| Benicar | EP0503785 | 0 | 1 | 0 | 0 | | 5 | |
| Bextra | WO1996025405 | 0 | 0 | 1 | 1 | | | 1 |
| Celebrex | WO1995015316 | 0 | 0 | 1 | 1 | | | 2 |
| Cialis | WO1995019978 | 1 | 1 | 0 | 1 | 1 | 0 | |
| Comtan | US4963590 | 1 | 0 | 0 | | 91 | | |
| Coreg | DE2815926 | 1 | 1 | 1 | 1 | 1 | 0 | 4 |
| Cozaar | EP0253310 | 1 | 0 | 0 | 0 | 7 | | |
| Diovan | EP0443983 | 1 | 0 | 1 | 1 | 1 | | 3 |
| Effexor | EP0112669 | 0 | 1 | 1 | | | 0 | 4 |
| Femara | EP0236940 | 1 | 1 | 0 | 0 | 32 | 5 | |
| Flovent | NL8100707 | 1 | 0 | 0 | 1 | 7 | | |
| Hycamtin | EP0321122 | 1 | 0 | 1 | | 1 | | 2 |
| Lamisil | EP24587 | 1 | 0 | 0 | 1 | 3 | | |
| Lansoprazole | EP0174726 | 1 | 1 | 0 | | 3 | 0 | |
| Lescol | WO1984002131 | 1 | 0 | 1 | 1 | 3 | | 2 |
| Paxil | US3912743 | 1 | 0 | 0 | 0 | 11 | | |
| Prandin | WO1993000337 | 1 | 1 | 1 | | 2 | 1 | 2 |
| Prograf | EP0184162 | 0 | 1 | 1 | | | 0 | 3 |
| Raxar | WO1989006649 | 0 | 0 | 1 | | | | 5 |
| Rezulin | EP0139421 | 1 | 0 | 1 | | 1 | | 3 |
| Spiriva | EP0418716 | 1 | 0 | 1 | 0 | 3 | | 3 |
| Starlix | EP0196222 | 1 | 1 | 1 | | 5 | 2 | 4 |
| Sustiva | EP582455 | 1 | 0 | 0 | 1 | 14 | | |
| Tarceva | WO1996030347 | 1 | 0 | 1 | 1 | 1 | | 5 |
| Viagra | EP0463756 | 1 | 0 | 1 | | 1 | | 1 |
| Vioxx | WO1995000501 | 0 | 1 | 1 | | | 0 | 1 |
| Xalatan | WO1990002553 | 0 | 1 | 1 | | | 7 | 4 |
| Zofran | DE3502508 | 1 | 1 | 1 | 1 | 1 | 0 | 2 |
| Zomig | WO1991018897 | 1 | 0 | 0 | 1 | 2 | | |
| Zyflo | EP0279263 | 1 | 1 | 1 | | 1 | 0 | 1 |

[a]A "1" stands for a successfully predicted key compound under the first five ranks; an empty field denotes that this patent specification was not contained in the prediction set. [b]An empty field denotes that this patent specification was not contained in the prediction set or the key compound was not predicted within the first five compounds in contrast to the first five ranks.

for this is clear. The largest maximum common substructure in the set of compounds retrieved from GVKBIO for the patent specification WO1991014679 includes an *ortho*-carboxylic acid functionality (Figure 3a) for 26 out of the total of 52 compounds. Avapro itself includes an *ortho*-tetrazole moiety (Figure 3b) in the corresponding position present in only five other exemplified compounds. The consequence is that Avapro is not in the most densely populated cluster and, hence, not predicted as a key compound. In this context, it should be mentioned that CSA also fails, whereas FOG successfully identifies Avapro as a key compound. The automatic series detection of the FOG method ranks a biphenyl scaffold on the first place (with 34 members), followed by the Avapro *ortho*-tetrazole biphenyl scaffold as shown in Figure 3b (with six members). Only the two compounds with the highest FOG-score per scaffold are considered, and Avapro is the top-ranked compound for the second scaffold.

**Performance of the FOG Method.** The FOG method correctly identifies 17 key compounds out of the 48 cases. This approach depends critically on the automatic series detection based on molecular frameworks. R-groups often contain rings or ring systems that do not belong to the core but can readily be perceived as such. The cyclooxygenase-2 selective inhibitor Bextra (WO1996025405) illustrates this. It is correctly predicted by Molecular Idol as the first ranked key compound, while both CSA and FOG methods fail. FOG's automated series detection algorithm produces eight scaffolds, and a pyridine derivative, rather than Bextra, is predicted with the highest FOG-score (Figure 4).

Defining the scaffold of interest using the Markush structure representation in the patent specification may enhance the prediction significantly.[32] Using the Markush scaffold shown in Figure 4c given in the document as input for the FOG method, 60 out of the 74 compounds are matched and Bextra is predicted as the first key compound.

In our experience the automatic series detection gives a useful overview of the patent specification's content, but as the prediction accuracy indicates, it does not perform as well as the two other methods (Table 1).

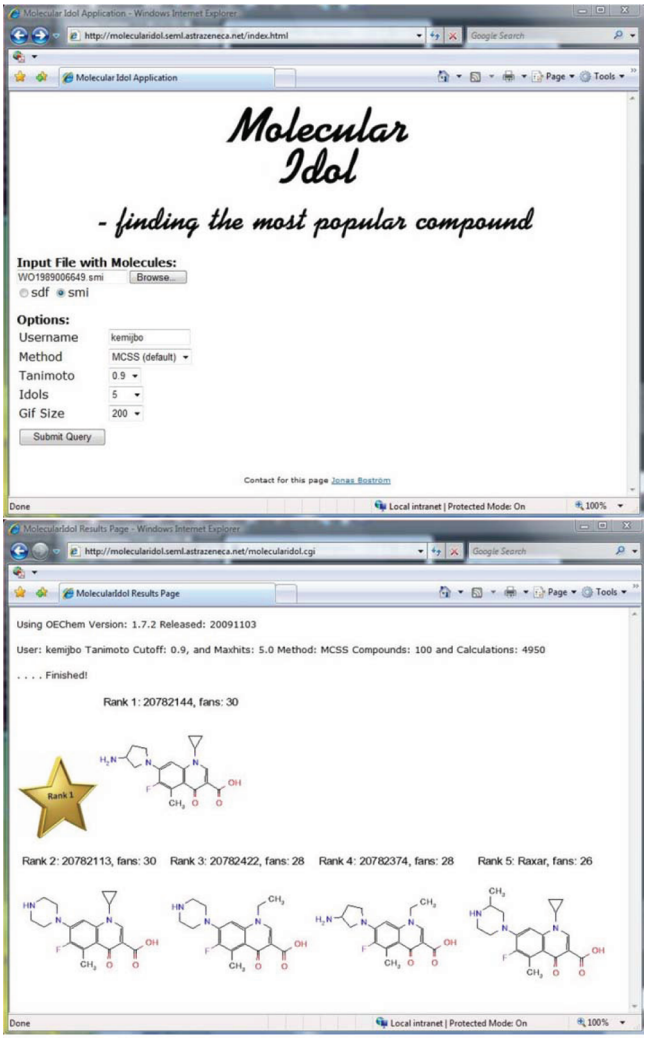



**Figure 2.** The Molecular Idol interface (top) and result page (bottom). Raxar is predicted among the first five key compounds (Rank 5).

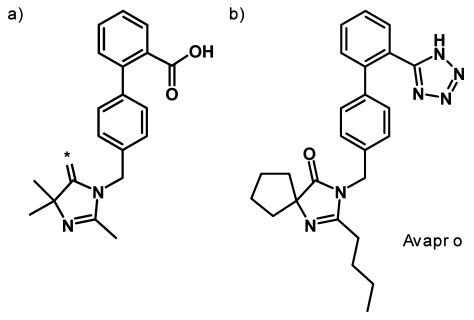



**Figure 3.** (a) The largest maximum common substructure for all compounds in WO1991014679 includes a carboxylic acid functionality. (b) The key compound Avapro includes an *ortho*-tetrazole leading to that it is not among the most popular compounds (i.e., the compounds with highest number of neighbors).

**Comparison of the Three Methods.** CSA performs better than FOG and Molecular Idol in terms of the number of correctly predicted key compounds. Nevertheless, the quality of prediction in terms of ranking order is better for the FOG method, whereas Molecular Idol retrieves the key compound for all correctly predicted drugs within the first five compounds

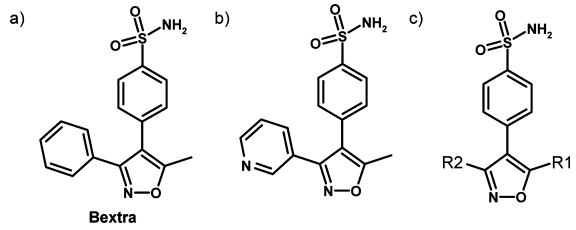



**Figure 4.** FOG analysis of patent specification WO1996025405: (a) Bextra, (b) the highest ranked key compound includes a pyridine instead of a phenyl, (c) the scaffold taken from a claimed Markush structure in the document which, if used as a starting point for FOG, will correctly identify Bextra as the key compound.

(Table 1). The reason for this behavior remains unclear. One can speculate that the MCS method of Molecular Idol is better at approximating diversity and, as such, captures the relevant scaffolds of the compound set. This could lead to the observed stricter ranking of compounds compared to that of a fingerprint method (Table 2). As mentioned above, the FOG method depends on the automatic series detection to identify the relevant scaffolds as well as on the composition of the compound set. If rings containing R-groups are used, which are not part of the key compound scaffold or the key compound is part of a less represented scaffold, this approach will fail.

Overall, 13 drugs are not within the top five ranked key compounds by any of the methods described here. We have attempted to identify the possible reasons for this by examining the first ranked key compound. In seven cases (Casodex, Livostin, Nasonex, Rescriptor, Rescula, Vigamox, and Zithromax), all methods predict the same incorrect compound on the first rank (Figure 5). For the remaining predictions (Azopt,

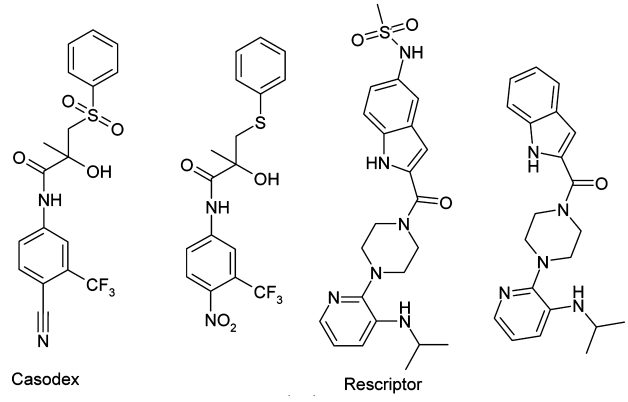



**Figure 5.** The top-ranked predicted key compound for the Casodex and Rescriptor patent specifications in all three methods is shown on the right side of the drug. As shown, the key compounds are structurally similar and mostly differ only in the scaffold decoration.

Camptosar, Detrol, Levitra, Omeprazole, and Reyataz), close analogues, with only few atom differences compared to the drug, are found on the highest rank (Figure 6).

No common underlying cause of prediction failures could be identified. This clearly exemplifies the difficulty in predicting key compounds without insights into the invention. Nevertheless, the correct scaffold is identified by all methods and can guide the analysis.

**Does Consensus Scoring Work?** From the 35 drugs (prediction accuracy 73%) correctly identified in their first


dx.doi.org/10.1021/ci3001293 | J. Chem. Inf. Model. 2012, 52, 1480−1489

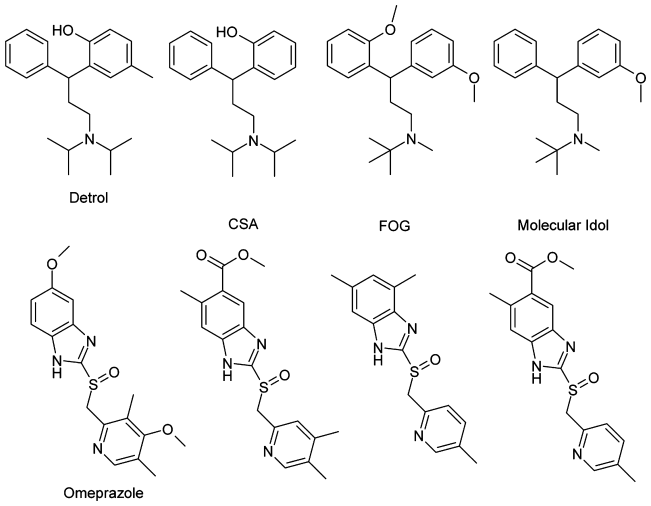



**Figure 6.** The top-ranked compounds for the prediction of Detrol and Omeprazole from their patent specifications in all three methods (CSA, FOG and Molecular Idol) are shown on the right side of the drug. As shown, the key compounds are structurally similar and differ only in the scaffold decoration.

patent specifications by the different approaches, eight (prediction accuracy 17%) are predicted by all three methods and an additional 14 by two out of three (Figure 7).

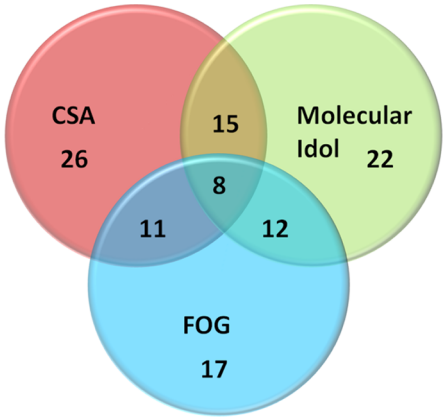
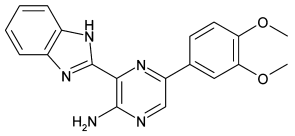



**Figure 7.** Venn diagram with the number of correctly predicted drugs by the three different approaches.

This indicates that the assumptions behind the different methods cannot be generalized and that the methods should be seen as complementary. Despite a thorough investigation, we could not identify a straightforward way to combine these methods to increase the overall prediction accuracy.

Another potential issue, inherent to all three methods, is that several compounds can have the same rank because of the same number of neighbors. Even if the cluster seed method has the highest prediction accuracy, it shows the widest spread in terms of number of predicted compounds. The average number of compounds needed to identify the key compound is 7.7 with a standard deviation of 17.8 in the case of CSA (due to Comtan with 99 compounds and Femara with 45 compounds, both in GVKBIO), whereas FOG and Molecular Idol predictions are much tighter (Table 2).

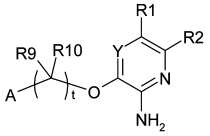
**Application of FOG Analysis in a Medicinal Chemistry Project.** A high throughput screen of the AstraZeneca compound collection identified a singleton hit pyridazin-2-amine compound **1** (Figure 8) as an ATP-competitive AXL



**1**

**Figure 8.** Singleton hit against ATP-competitive AXL kinase identified in HTS of AstraZeneca collection.

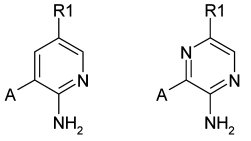
kinase inhibitor, which was potent (Table 4) but suffered from low solubility and poor physicochemical properties (Table 4). At a similar time, three patent applications were published by Pfizer claiming pyridin-2-amines and pyridazin-2-amines as inhibitors of cMet kinase, a close neighbor of AXL (Figure 9).



| Patent | Claims | No. of examples in GVKBIO database |
|---|---|---|
| WO2006021881 | R1 = 5-membered rings | 152 |
| WO2006021884 | A = chiral tri-halotoluyl | 19 |
| WO2006021886 | R1 = 6-membered rings | 155 |

**Figure 9.** Patent applications published by Pfizer in 2006.

The structural similarity between the two kinase targets and the hinge binding motifs of the inhibitors intrigued us, and FOG analysis was performed on the three documents in an attempt to identify the key compounds in each. Therefore, the exemplified compounds from each patent specification were extracted from the GVKBIO database, and the two cores in Figure 10 were used as input for the FOG analysis.



Pyridin-2-amine core Pyridazin-2-amine core

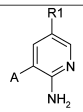
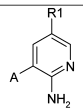
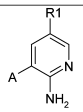
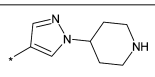
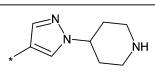
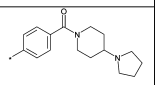
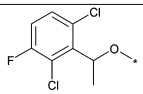
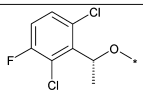
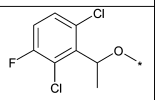
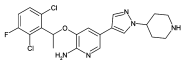
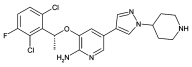
**Figure 10.** Core structures used in fragmentation routine for FOG analysis.

The key compounds predicted by FOG for the three documents are presented in Table 3.

A graphical visualization of FOG analysis for WO2006021881 is presented in Figure 11. Each point represents an exemplified compound: blue points represent compounds with the pyridine-2-amine core and red points contain the pyrazin-2-amine core. Rows and columns of points represent the side chains that are used most frequently in compound examples. The points at which these sets of compounds cross typically represent the key compounds because these are combinations of frequently used groups. In Figure 11, the cross point of the two most heavily populated series is example 30 in the patent specification.


dx.doi.org/10.1021/ci3001293 | J. Chem. Inf. Model. 2012, 52, 1480−1489

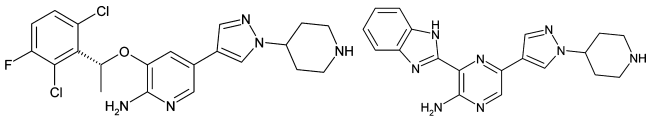
**Table 3. FOG Analysis Results of Three Pfizer Patent Specifications**

| | WO2006021881 | WO2006021884 | WO2006021886 |
|---|---|---|---|
| Favourite Core |  |  |  |
| Favourite R1 group |  |  |  |
| Favourite A group |  |  |  |
| Predicted best mode compound |  Example 30 |  Example 13 | Combination not actually an example in this document – see WO2004076412, example i-294 |

First, FOG analysis of WO2006021886 revealed that the exemplified compounds contained numerous examples of infrequently used R1 groups. Furthermore, the combination of the two most frequently used R-groups for the A and R1 group (Figures 9 and 11) was not claimed in the patent application but had been previously described in an earlier patent specification. These observations lead us to the conclusion that WO2006021886 is more likely an extension to widen the patented scope rather than a patent specification containing a clinical candidate. Therefore, we focused our further analysis on WO2006021881 and WO2006021884.[33]

Applying the FOG methodology revealed that the 1-(2,6-dichloro-3-fluoro-phenyl)ethoxy group was consistently the preferred R-group in all three documents, albeit the chiral (R) form in WO2006021884. On the basis of previous knowledge, we knew that the benzylether A group was particularly favorable for cMet kinase, leading our attention toward the R1 group. The 1-(4-piperidyl)pyrazol-4-yl R1 group, which was highlighted by the FOG approach (i.e., a preferred R1-group in the two patent specifications), occupies the space leading toward solvent,[34] and the low lipophilicity of the side chain compared with the 3,4-dimethoxyphenyl group in **1** was considered a favorable property.

It was decided to synthesize compound **3**, fusing the 1-(4-piperidyl)pyrazol-4-yl group from **2** with the alternative hinge binder of **1** (Figure 12). Biological testing confirmed that AXL kinase tolerated this change, and potency was retained.



**2** **3**

**Figure 12.** Key compound **2** identified from cMet patent applications and hybrid with HTS hit **3**.

Furthermore, the introduction of the 1-(4-piperidyl)pyrazol-4-yl side chain reduced lipophilicity of the molecule significantly, which in turn improved physicochemical properties of the compound (Table 4).[35]

**Table 4. Improved Physicochemical Properties for Compound 1 and 3[35]**

| | Compound **1** | Compound **3** |
|---|---|---|
| AXL enzyme $pIC_{50}$ | 8.0 | 8.3 |
| AXL cell $pIC_{50}$ | 7.3 | 6.9 |
| cLogP | 3.5 | 1.5 |
| logD (Shake flask pH 7.4) | too low solubility | 1.8 |
| Solubility at pH 7.4, $\mu$M | <0.5 | 31 |
| Rat plasma protein binding | >99% | 93% |

Alongside the design of compound **3** guided by the FOG approach, analogues of **1** were systematically synthesized as a compound library using available boronic acid reagents to replace the R1 R-group. Around 50 compounds were made to extend the known SAR as to explore the property space with respect to lipophilicity. None of library compounds brought benefits as significant as the hybrid compound **3**, suggesting that the FOG analysis can easily and quickly provide significant information about compound optimization possibilities. In this case, we identified the unusual and noncommercially available 1-(4-piperidyl)pyrazol-4-yl group, which effectively reduced project time. Compound **3** fulfilled our internal lead generation criteria and allowed the project to progress into a later lead optimization stage; the results of which will hopefully be published in due course.

## CONCLUSIONS

The methods described in this work represent convenient and efficient ways for the medicinal chemist to gain an understanding of the structural coverage of a patent specification.

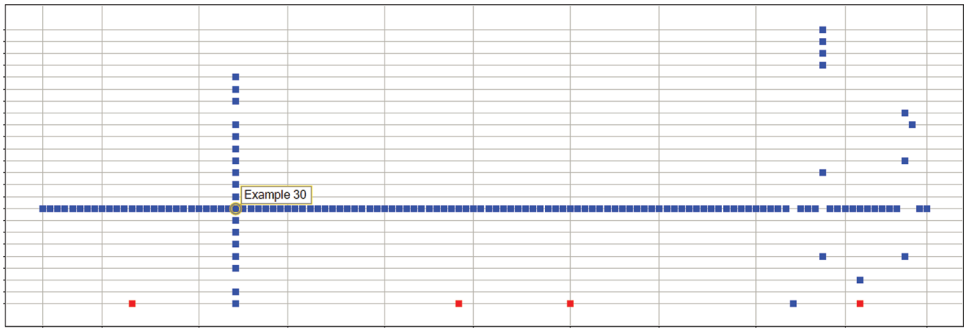



**Figure 11.** Graphical view of WO2006021881 exemplified compounds; identity of R1 ($x$-axis) plotted against A ($y$-axis).


dx.doi.org/10.1021/ci3001293 | J. Chem. Inf. Model. 2012, 52, 1480−1489

This eases the work of the medicinal chemistry expert who still needs to judge the predictions and select key compounds for further evaluation. The prediction accuracy of CSA exceeds that of FOG and Molecular Idol, but the latter two offer a more robust enrichment and a better visualization of the scaffolds covered in the patent specification. By manual selection of a scaffold based on the Markush structure claimed in the document, the accuracy of the FOG prediction may improve.

Our results indicate that structural information alone is most likely not enough in terms of a successful prediction of a key compound from a patent specification. This may be partially due to the structural data used as a basis for prediction as different data providers have different business rules for patent chemistry extraction. Another reason is that clinical candidates are defined by properties such as toxicity or PK/PD profiles. These properties are complex, and it is hard to predict endpoints that are linked to the structure often in a nonobvious way. Patent specifications are documents written to protect the invented drugs in court and not to describe the SAR or the chemistry development.

In future work, we plan to include additional information such as PK/PD data, as well as investigate the use of calculated properties to further enhance key compound predictions.

## ■ ASSOCIATED CONTENT

### Supporting Information

The drug name, the drug structure (SMILES), the relevant patent specification, and number of compounds in the different data sources. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: christian.tyrchan@astrazeneca.com.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Lucy Padget for the valuable assistance with intellectual property terminology.

## ■ REFERENCES


(1) Hattori, K.; Wakabayashi, H.; Tamaki, K. Predicting key example compounds in competitors' patent applications using structural information alone. *J. Chem. Inf. Model.* **2008**, *48*, 135−142.

(2) Engel, T. Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267−2277.

(3) Southan, C.; Varkonyi, P.; Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminf.* [Online] **2009**, *1*, Article 10. http://www.jcheminf.com/content/1/1/10 (accessed October 31, 2011).

(4) Bachrach, S. Chemistry publication - making the revolution. *J. Cheminf.* [Online] **2009**, *1*, Article 2. http://www.jcheminf.com/content/1/1/2 (accessed October 31, 2011).

(5) DiMasi, J. A.; Faden, L. B. Competitiveness in follow-on drug R&D: A race or imitation? *Nat. Rev. Drug Discovery* **2011**, *10*, 23−27.

(6) Trippe, A. J. Patinformatics: Tasks to tools. *World Pat. Inf.* **2003**, *25*, 211−221.

(7) Bonino, D.; Ciaramella, A.; Corno, F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Pat. Inf.* **2010**, *32*, 30−38.

(8) Moehrle, M. G.; Walter, L.; Bergmann, I.; Bobe, S.; Skrzipale, S. Patinformatics as a business process: A guideline through patent research tasks and tools. *World Pat. Inf.* **2010**, *32*, 291−299.

(9) Dou, H. J.-M. Benchmarking R&D and companies through patent analysis using free databases and special software: A tool to improve innovative thinking. *World Pat. Inf.* **2004**, *26*, 297−309.

(10) Fattori, M.; Pedrazzi, G.; Turra, R. Text mining applied to patent mapping: A practical business case. *World Pat. Inf.* **2003**, *25*, 335−342.

(11) Kim, Y. G.; Suh, J. H.; Park, S. C. Visualization of patent analysis for emerging technology. *Expert Syst. Appl.* **2008**, *34*, 1804−1812.

(12) SciFinder. http://www.cas.org/products/sfacad/index.html (accessed October 31, 2011); STN. http://www.cas.org/products/stnfamily/index.html (accessed October 31, 2011).

(13) Kaback, S. M. What is in a patent? Information: But can I find it? *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 159−163.

(14) Kaback, S. M. A patent searcher's personal chronicle: 40 years in the evolution of a profession. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 137−142.

(15) Banville, D. L. Mining chemical structural information from the drug literature. *Drug Discovery Today* **2006**, *11*, 35−42.

(16) CAS databases. http://www.cas.org/expertise/cascontent/index.html (accessed October 31, 2011).

(17) Thomson Reuters Pharma. http://thomsonreuters.com/products_services/science/science_products/a-z/thomson_pharma/ (accessed October 31, 2011).

(18) *GOSTAR*. http://www.gostardb.com/gostar/loginEntry.do (accessed October 31, 2011).

(19) SureChem. http://www.surechem.org (accessed October 31, 2011).

(20) Spangler, S.; Ying, C.; Kreulen, J.; Boyer, S.; Griffin, T.; Alba, A.; Kato, L.; Lelescu, A.; Yan, S. Exploratory analytics on patent data sets using the SIMPLE platform. *World Pat. Inf.* **2011**, *33*, 328−339.

(21) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making every SAR point count: The development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* **2011**, *16*, 1019−1030.

(22) Lepp, Z.; Huang, C.; Okada, T. Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. *J. Chem. Inf. Model.* **2009**, *49*, 2429−2443.

(23) European Patent Convention. Article 100. http://www.epo.org/law-practice/legal-texts/html/epc/2010/e/ar100.html (accessed October 31, 2011); Article 138. http://www.epo.org/law-practice/legal-texts/html/epc/2010/e/ar138.html (accessed October 31, 2011).

(24) Example of patent invalidated by the US courts for failure to comply with the best mode requirement. http://www.cafc.uscourts.gov/images/stories/opinions-orders/09-1081.pdf (accessed October 31, 2011).

(25) GVKBIO Drug Database. http://www.gvkbio.com/database_pdf/Drug.pdf (accessed October 31, 2011).

(26) Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/ (accessed October 31, 2011).

(27) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49*, 6716−6725.

(28) Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. MultiMCS: A fast algorithm for the maximum common substructure problem on multiple molecules. *J. Chem. Inf. Model.* **2011**, *51*, 788−806.

(29) OEChem TK. http://www.eyesopen.com/oechem-tk (accessed October 31, 2011).

(30) Rabal, O.; Urbano-Cuadrado, M.; Oyarzabal, J. Computational medicinal chemistry in fragment-based drug discovery: What, how and when. *Future Med. Chem.* **2011**, *3*, 95−134.

(31) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(32) Simmons, E. S. Markush structure searching over the years. *World Pat. Inf.* **2003**, *25*, 195−202.

(33) With the benefit of hindsight, it is now apparent that example 13 in WO2006021884, which was identified as the best mode compound by FOG analysis, is in fact crizotinib (Xalkori), Pfizer's anaplastic lymphoma kinase/c-Met inhibitor, which was launched as a treatment for non-small cell lung carcinoma in 2011.

(34) Cui, J. J.; Tran-Dubé, M.; Shen, H.; Nambu, M.; Kung, P.-P.; Pairish, M.; Jia, L.; Meng, J.; Funk, L.; Botrous, I.; McTigue, M.; Grodsky, N.; Ryan, K.; Padrique, E.; Alton, G.; Timofeevski, S.; Yamazaki, S.; Li, Q.; Zou, H.; Christensen, J.; Mroczkowski, B.; Bender, S.; Kania, R. S.; Edwards, M. P. Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal−epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). *J. Med. Chem.* **2011**, *54*, 6342−6363.

(35) Synthesis and biological testing information for compound 3 has been reported in patent WO2009024825.