Journal of
CHEMICAL INFORMATION
AND MODELING

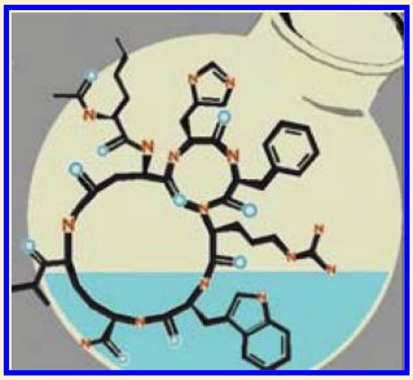
Article

pubs.acs.org/jcim

AsteriX: A Web Server To Automatically Extract Ligand Coordinates from Figures in PDF Articles

V. Lounnas* and G. Vriend

CMBI NCMLS Radboud University, Nijmegen Medical Centre, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands



ABSTRACT: Coordinates describing the chemical structures of small molecules that are potential ligands for pharmaceutical targets are used at many stages of the drug design process. The coordinates of the vast majority of ligands can be obtained from either publicly accessible or commercial databases. However, interesting ligands sometimes are only available from the scientific literature, in which case their coordinates need to be reconstructed manually—a process that consists of a series of time-consuming steps. We present a Web server that helps reconstruct the three-dimensional (3D) coordinates of ligands for which a two-dimensional (2D) picture is available in a PDF file. The software, called AsteriX, analyses every picture contained in the PDF file and attempts to determine automatically whether or not it contains ligands. Areas in pictures that may contain molecular structures are processed to extract connectivity and atom type information that allow coordinates to be subsequently reconstructed. The AsteriX Web server was tested on a series of articles containing a large diversity in graphical representations. In total, 88% of 3249 ligand structures present in the test set were identified as chemical diagrams. Of these, about half were interpreted correctly as 3D structures, and a further one-third required only minor manual corrections. It is principally impossible to always correctly reconstruct 3D coordinates from pictures because there are many different protocols for drawing a 2D image of a ligand, but more importantly a wide variety of semantic annotations are possible. The AsteriX Web server therefore includes facilities that allow the users to augment partial or partially correct 3D reconstructions. All 3D reconstructions are submitted, checked, and corrected by the users domain at the server and are freely available for everybody. The coordinates of the reconstructed ligands are made available in a series of formats commonly used in drug design research. The AsteriX Web server is freely available at http://swift.cmbi.ru.nl/bitmapb/.

INTRODUCTION

Over the years, millions of small molecules have been synthesized and have been published in scientific articles describing their relevance for the pharmaceutical industry in terms of binding to receptors, drug-likeness, accessibility in the body, etc. Many of these small molecules have found their way into small molecule databases or information systems like ChemBank,[1] ChemPDB,[2] KEGG,[3] or TimTec.[4] PubChem[5] probably provides the largest collection of low molecular weight compounds. The ChEMBL database[6] has collected more than 600,000 publicly accessible small molecules with literature references attached to them.

Over the years, large collections of compounds have found wide use in in vivo and in vitro high-throughput screening in pharmaceutical research. Recently, focus is shifting toward screening with small, focused libraries that often revolve around novel molecules recently designed in academia and that often have not found their way into the databases yet. In those cases one can use software like the JME ligand editor[7] to draw the molecule, convert it to a SMILES or InChI representation, and use software such as Corina[8] or MolConverter from ChemAxon[9] to generate 3D coordinates. Although very doable, this is a tedious process, especially when a series of large ligands has to be extracted from an article.

CLiDE,[10] its recently upgraded version,[11] and OSRA[12] can automatically extract ligand structures from 2D graphical representations such as those usually appearing in patent spreadsheets. Patent spreadsheet ligands are represented minimally by a set of vectors (bonds) and characters (explicit heteroatoms) that are connected to each other unambiguously to avoid confusion. Patents are written to be optimally unambiguous to avoid any possible legal problems. In many graphical representations of ligands in the scientific literature, however, a large number of complications occur, especially in recent years with the advent of more sophisticated drawing softwares.

The first step in processing images other than patent spreadsheets is to automatically recognize and separate the areas in the images that may contain representations of small molecular structures. There is a need for image processing methods allowing the ligand containing areas to be delineated before the structures are actually solved in details. Even when ligand-containing areas can be correctly identified, a whole array of complications remain that, although easy to resolve by visual inspection, represent a considerable challenge for a computer. For instance, this occurs when elements of the drawing that should be distinct actually overlap each other, such as segments representing bonds and letters representing atom labels or in the cases of cage-compounds where segments

Received: September 12, 2011
Published: February 2, 2012

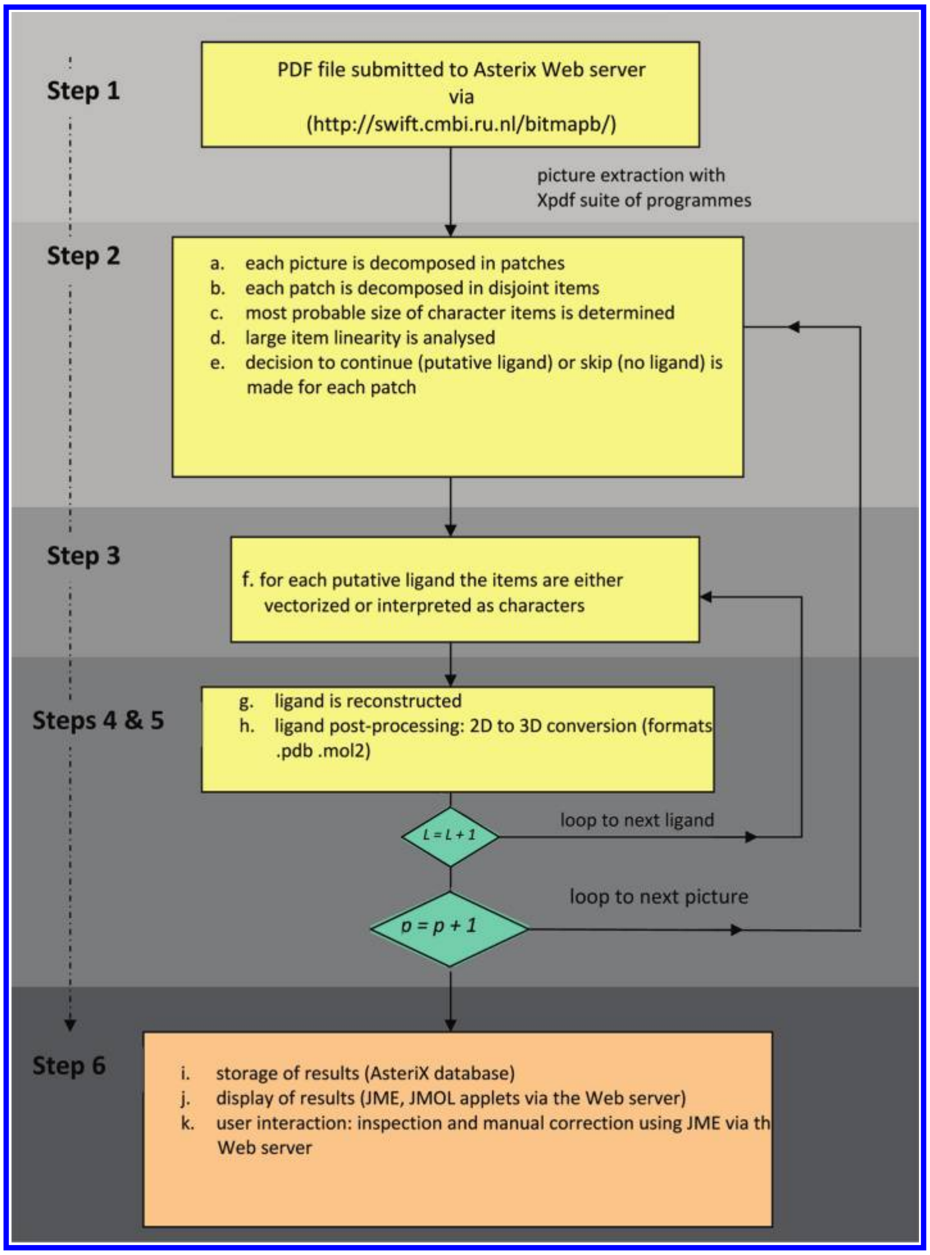

Step 1
PDF file submitted to Asterix Web server
via
(http://swift.cmbi.ru.nl/bitmapb/)
picture extraction with
Xpdf suite of programmes
Step 2
a. each picture is decomposed in patches
b. each patch is decomposed in disjoint items
c. most probable size of character items is determined
d. large item linearity is analysed
e. decision to continue (putative ligand) or skip (no ligand) is made for each patch
Step 3
f. for each putative ligand the items are either vectorized or interpreted as characters
Steps 4 & 5
g. ligand is reconstructed
h. ligand post-processing: 2D to 3D conversion (formats .pdb .mol2)
L = L + 1
loop to next ligand
p = p + 1
loop to next picture
Step 6
i. storage of results (AsteriX database)
j. display of results (JME, JMOL applets via the Web server)
k. user interaction: inspection and manual correction using JME via th Web server

Figure 1. AsteriX flowchart. Diamond boxes and arrows indicate the hierarchy of operations. AsteriX loops over all pictures in the PDF ($P = P + 1$) and per picture over all small molecules it contains ($L = L + 1$).

representing separate bonds cross each others. The situation becomes even more difficult when the ligand representation differs from the prototypical description found in patent spreadsheets. It may contain additional elements such as brackets, substitution pointers, wiggly bonds, or semantically loaded artifacts such as curved arrows that interrupt or overlap segments representing bonds to indicate a reaction centers or a bond rearrangement. The AsteriX software and Web server were designed to address the broader context of published scientific articles in which ligand representations often are nonstandard but either artistically adapted or cluttered with annotations or combined with other graphical objects.

■ METHODS

The AsteriX process consists of six distinct steps (Figure 1) that start with PDF reading and decomposition and end with the generation and optional manual correction of 3D ligand coordinates. These six steps are as follows:

1. Read and decompose PDF files
2. Detect pictures that contain representations of small molecules
3. Analyze small molecules in terms of characters (any cipher or letter that may represent an atom label or that may be part of an annotation or the short hand notation of a chemical substituent) and bonds
4. Combine bonds and atom names into a reconstructed topology
5. Use MolConverter (of ChemAxon) to generate 3D coordinates
6. Optional user interaction

These six points will be elaborated in the next six paragraphs. The number of problems (and corresponding solutions) is so large that a complete enumeration is way beyond the scope of this journal. Interested readers can obtain the documented source code to get access to many more than the few (key) algorithmic components listed here.

Step 1. Read and Decompose PDF Files. Pictures and texts are extracted from PDF files using the Xpdf suite of programmes for Linux.[13] Journal names and issue specific information (volume, pages, year) are identified using a simple character matching algorithm that compares the first 25 records extracted from the PDF with a precompiled dictionary of journal names. Issue specific information (volume, pages, years) are extracted and stored in the database, too.

The Xpdf PDF extracted pictures are either in .ppm or .pbm binary format and converted to .ppm ASCII formatted files that are stored for further processing. Color pictures are converted to black and white. A cutoff of 50% saturation is used to decide whether a pixel will be white (background) or black (putative ligand).

Picture dimensions up to 12000 × 12000 pixels can be dealt with. When the picture width exceeds 1800 pixels, resizing is performed to avoid excessive CPU time.

To avoid legal issues, Asterix does not maintain PDF files in its database but only the bibliographic information and essential components of the ligand containing pictures in a proprietary format.

Step 2. Detect Pictures That Contain Representations of Small Molecules. Many pictures in PDFs contain multiple ligands that have to be processed separately. Sometimes the ligands are annotated by graphical objects such as a protein ribbon or a variety of arrows, curves, or diagrams. In addition, frames can be present around the ligands. Ligand representations often are discontinuous because of minor drawing or printing imperfections or because lines were shortened to make space for text labels. Detection of individual ligands in a picture is therefore done in AsteriX in a two step process. First, all pixels that are within a 19 × 19 pixel box centered on an originally black pixel are made black to ensure that small discontinuities are bridged. Contiguous black pixels are then collected in groups and, after going back to the original pixels, stored as likely ligands that we call "patches". The patches are subsequently broken up into their constituting "motifs" (lines and characters). A motif is defined by a subset of black pixels that are all connected to each other and that represent either bonds or atom types/atom names.

Seven criteria (a–g) are at the basis of the decision for whether a patch contains a likely ligand that must be analyzed further or should be discarded:

a. Kill patches with less than 50 pixels along both the *X* and the *Y* axis
b. Kill patches with less than 10 or more than 120 motifs
c. No vertical or horizontal vector is allowed longer than 70% of the patch dimensions
d. No contiguous straight line in either the *X* or *Y* direction may exceed 200 pixels
e. At least one character must be found representing either one of the atom types C, N, O, H, S, P, or F, or one of the commonly used substitution labels R, X, Y, M, or W. This criterion seems to unnecessarily eliminate some ligands. Experience tells us, however, that all interesting structures contain noncarbon atoms that are labeled with a character, while very few carbon-only structures like anthracene are found in the biomedical literature. On the other hand, many line drawings do exist. Figure 2 illustrates this topic.

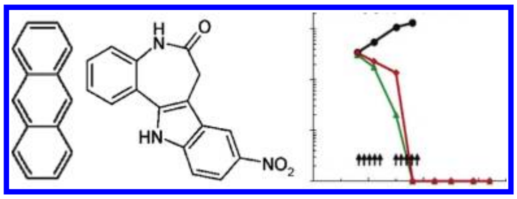


Figure 2. Anthracene, left, does not require any characters. Still it is excluded as such character-free ligands are very rare, while nonligand pictures containing character-free areas (patches) like the one drawn at the right-hand side[14] are very abundant. Medicinally interesting structures, like the one in the middle,[15] almost invariably contain atoms other than carbon that have been indicated by a character label.

R, X, Y, W, and M are accepted as labels because they are widely used to indicate the presence of either groups of hetero atoms or series of substitution groups. R and S may also indicate enantiomeric states.

The average character size is a crucial parameter to determine well, as it is needed to resolve problems arising when pixels are shared by characters representing atoms and lines representing bonds. Characters are interpreted by convoluting their motif with a collection of prestored characters for which pixels are set to black such that the most commonly used font types are covered. Once a character is detected, its *X* and *Y* dimensions and location are stored, and all pixels that belong to that character are switched from black to white. Special code deals with problematic cases when characters and bonds share pixels (see Step 3).

f. Upon determining if a patch is a putative ligand or some nonligand drawing, all possible straight line segments (11 pixels long) through continuous patches of black pixels are determined. One single bond can at this stage easily be represented by a few dozen straight lines, especially if the lines in the drawing are multiple pixels wide. These line segments should account for 85% of the black pixels composing the noncharacter motifs in a patch. This is what we refer to as "linearity".
g. When bonds come together in 2D pictures, they tend to do so using a limited number of angles (0, 60, 90, but mostly 120 degree; Figure 3). We observed that normally at least 55% of all line segments meet at such "structure-like" angles.

A priori, one would expect that there are two characteristics that are different between drawings of ligands and all other drawings. First, the length distribution of lines in a drawing of a ligand should be very narrow, and second, the distribution of angles between those lines should spike around specific values such as 0, 60, 90, 120 degrees. In practice, bond length distributions tend to be much more variable than expected, and therefore, only the angle distribution can be used to infer "structure likeness".

Out-of-plane bonds normally are represented by either a solid or a dashed triangle. They are detected with special pattern recognition procedures. The identification of solid triangles is done either as part of the detection of characters or during the vectorization process, depending on its relation to neighboring bonds and characters. Dashed triangles are hardest

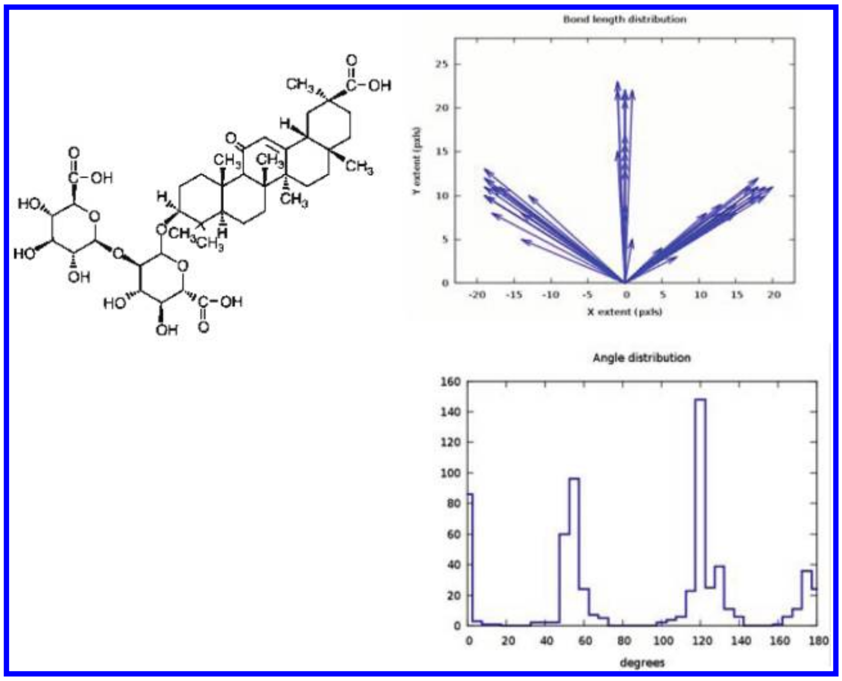


Figure 3. Top left: Structure of, glycyrrhizic acid, a liquorice derivative.[16] Top right: Direction vectors for all lines observed in glycyrrhizic acid. Vectors are draw in the positive *Y* direction. Bottom right: Histogram of angles between the bond-lines in glycyrrhizic acid.

to detect and are therefore identified further down the analysis process when all other options ran out

Step 3. Analyze Each Putative Small Molecule in Terms of Characters and Line Segments or Other Representations of Chemical Bonds. We consider a ligand drawing as an irregular polyhedral object or connected graph projected in a two-dimensional plane.[17,18] In general, the pixel density along the vertices representing the bonds is lower in the middle than at their extremities where they connect to other vertices (bonds) or alphabetical characters (atom labels). This property is sufficiently robust to be exploited in many different ligand representations used in the scientific literature. This principle is also at the basis of AsteriX′ algorithm to detect overlap between atom labels and bonds. This algorithm calculates the density around each black pixel and on the basis of the distribution of values provides an estimate of what may actually be the middle or the end of a bond where there tend to be other things present like character labels, triangles, etc. The threshold is evaluated during the process for each analyzed patch using the density distribution values. Once these regions are identified, the algorithm makes use of an estimate of the atom label dimensions (height and width; see Step 2) to separate putative characters from other semantic enrichments. Figure 4 shows the higher density regions detected in the drawing of a ligand where overlapping atom labels could be identified and separated using this approach.

Once it is clear that a patch contains characters and bonds, their nature must be analyzed in detail. For all lines, the endpoints must be determined. Endpoints of lines must be matched. Characters must be associated with endpoints of lines as those endpoints are at or near the positions of atoms. Characters that do not match endpoints of lines are stored to be reassessed at a later stage.

A large series of problems must be solved in Step 3. For example, disconnected bonds must be connected. When characters and lines share pixels, only the character pixels that are not shared with the line must be set to white after the character was analyzed. Nonbond lines such as arrows that often are added to a drawing to increase the information content can make the interpretation difficult and sometimes even impossible. If a bond is broken to make space for a character, then the character must be removed, and the resulting lines must be extended to complete the bond again. Circles and arcs that represent other things than atoms or bonds must be detected and cleanly removed, etc. Figures 6 and 7 illustrate a few of these problems. Figure 6 shows cases that AsteriX dealt with properly, while Figure 7 shows a series

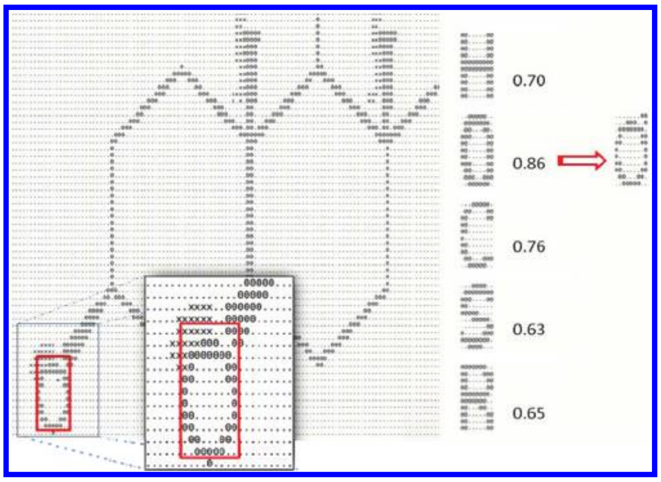


Figure 4. Left: Fragment of a typical ligand drawing. Small dots represent white pixels, large dots black ones. Sets of small crosses indicate the top left corners of boxes (of the dimension of a character) in which high density characteristics were determined. The red box at the bottom left has been drawn by hand for clarity. This box, like each box, is convoluted with a large series of pixel patterns for characters. The patterns for the characters H, O, C, S, and R are shown (with the matching score for the red-boxed character indicated), and the content of the red box is repeated at the far right to indicate that the match with the best score (0.86) is with the pattern for O.

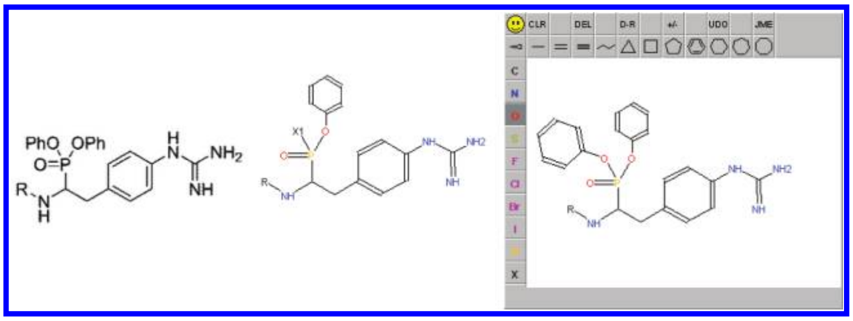


Figure 5. Typical example of required user interaction. Left: Bitmap of the molecule as extracted from the PDF.[22] Middle: AsteriX′ interpretation. One of the two phenoxy groups (OPh) is "missed" and labeled X1 to aid the user's interactive correction. The R-group could even indicate a series of substituents so that the user needs to make a decision what to replace it with. Right: JME window from the correction page in which the user corrected the PhO problem but left the R-group ambiguous.

of examples that are still beyond the artificial intelligence level of today's software.

Step 4. Combine Bonds and Atom Names into a Reconstructed Topology. The combination of bonds and atom names into a topology requires a series of steps. All characters, normal bonds, and out-of-plane bonds must be tabulated. Atoms connected by two or three lines must be marked as double or triple bonded. In cases where aromaticity is explicit indicated, it can be detected by comparing the position of the centers of the 5- and 6-member rings with observed circles and ellipses (benzene is ellipsoidal rather than round in many ligand pictures). Out-of-plane bonds must be properly connected to in-plane atoms. Special code is needed to deal with isolated vertical lines as those can either be bonds or the character i in either lower or upper case.

The final steps include dealing with a large number of small details, most of which are implemented rules of thumb. Any atom that is too far away from a character is made a carbon. Groups of closely spaced characters, like COOH, NH2, CH3, C2H5, OH, Br, etc., must be interpreted and converted into additional atoms and bonds. Molecular substituent that cannot be converted because they are not present in the AsteriX substituent dictionary are stored and referenced in the output SDF file of the ligand so that they can be reassessed later on, either by future improvements of the software or until that moment by the user. Branching points on the ligand scaffold are marked with an index that references them. Unassigned characters must be explained either as part of a two-letter chemical element such as Fe or Cu, kept as an atom annotation, or removed when deemed to be situated too far from any bond or atom position. For instance, stereochemical indicators such as R, S, D, or L will be stored as atom annotation and kept in the ligand structure output file (SDF format). Plus and minus signs are also converted into plausible atom charges (COO+, for example, is indicative of an error). Character strings like R1, R2, where R1 and R2 are later explained to be any of, for example, methyl, ethyl, phenyl, etc., must be stored to later indicate incompleteness of the ligand and to avoid, for example, placing carbons at such locations. These annotations aid the AsteriX Web server user in completing the ligand manually.

Throughout the process many other empirically adjusted cut-offs are used and scores determined. For instance, how far can an atom label be away from a bond extremity? How far can two lines of a double bond be separated? How far can two line endpoints be away from each other while still being connected to the same atom center? All these empirical cut-offs could in principle be related to probability scores for the quality of their individual determination. All these probabilities would then at the end be combined into a single score that indicates the estimated overall ligand determination quality score. At this moment, such a scoring scheme has not yet been implemented, albeit that provisions for this future extension are included.

Step 5. Use ChemAxon To Generate 3D Coordinates. The coordinates extracted from the picture analysis are converted into an SDF[19] file. This 2D structure data file contains the following:

a. A unique structure identification code calculated from the journal name, volume, issue, figure number, patch number, and the geometrical center of the patch relative to the origin of the whole figure
b. 2D coordinates of the structure
c. Electrostatic charges if any
d. Name of the compound if determined
e Journal parameters (name, volume, issue, page numbers)
f. Atom annotations
g. Unresolved chemical substituents

One SDF file is produced per ligand. The ChemAxon software[9] is used to calculate 3D coordinates from these SDF files. At this moment, the 3D structures are produced without hydrogen atoms, which means that the information on the protonation state of the titratable groups present in the SDF file is lost. The 3D coordinates are returned to the user in three formats: PDB format,[20] MOL2 format,[21] and 3D SDF format.

Step 6. User Interaction. Figure 6 clearly shows that it will for a long time remain impossible to convert all pictures into correct 3D coordinates. The AsteriX software therefore produces for each ligand the original digitized picture and its 2D interpretation in the same orientation and presents these two pictures side-by-side to the user who can then use the JME small molecule editor to insert missing groups, remove overinterpretations, and make corrections (see Figure 5 for a typical example). When the user is ready, the corrected 2D representation will be converted into 3D coordinates again, and the corresponding database entry can be overwritten with the new coordinates by a curator or any trusted party who received the curator password. All SDF and coordinate files are finally packed up in a compressed .tgz file that can be downloaded by the user.

■ RESULTS

In the process of writing AsteriX, we tested it on 167 articles in PDF format. Most of these articles were obtained from the *Journal of Medicinal Chemistry*[23] because nearly each article in this journal contains pictures of medicinally relevant ligands. The molecular representations in *J. Med. Chem.* are not standardized, so the picture styles and dimensions vary extremely throughout this test set. Eighty-eight percent (88%) of the 3249 ligand structures present in the test set could be detected. About half of all structures were interpreted correctly, while about 15% of the molecules were not converted well enough to even detect the overall shape. About a third of the structures required just a few minor manual corrections with JME. Figure 5 shows one example from this category. Typical minor errors include the following:

a. Missing bonds because the line was drawn too short or was hidden because of overlap with an explicit atom label or annotation
b. Presence of extra bonds resulting from wrongly interpreted arrows or annotations located very near bonds or atom labels
c. Atom label misinterpretation caused by the imperfection of the character recognition algorithm
d. Chemical substituents not converted to molecular substructures because they are either not present in the substituent dictionary of AsteriX or may have been misinterpreted by the character recognition algorithm (like in the example in Figure 5)

A large series of very heterogeneous reasons underlie the 15% of all cases in which AsteriX fails completely. These range from incomplete or partially truncated ligand structures due to too wide empty spaces, imperfectly/partly assessed connectivities, too many bonds that cross each other, an overload of semantic annotations, or highly variable bond lengths (see Figure 7 for a series of examples). The percentage of false positives (pictures patches that were incorrectly interpreted as ligands) is below 5%. These false positives are not really a problem as they normally lead to molecules with ridiculous chemistry, and the user can remove them with a single click.

We have used the MOLPRINT 2D software[24] to benchmark AsteriX on the OSRA test set that consists of 5735 ligand structures extracted from US patents. The MOLPRINT atom environment descriptors were modified to make aromatic carbon atoms equivalent to sp2 carbon and aromatic and amide nitrogens equivalent to sp2 nitrogens. Tanimoto indices where calculated with a depth of 1 atom layer. For this set, we obtain a Tanimoto index of 0.90 with AsteriX and 0.95 with OSRA. We have used the OSRA test set to detect and solve bugs in our algorithms, but we did not use it to optimize the accuracy of AsteriX. We have also computed the Tanimoto similarity index between the ligands as interpreted by AsteriX and the manually corrected ligands for a subset of 39 articles. This test set contains a series of ligands that hold representative recognition problems. For the 670 ligand structures extracted from these articles by AsteriX, we obtain an average Tanimoto index of 0.85 compared with 0.74 by OSRA. The Tanimoto index has the advantage of being objective, quantitative, and reproducible, but it does not always represent the intuitive evaluation of the quality of the interpretation. We believe that visual inspection can bring more quality appraisal than intricate molecular similarity index calculation. The aforementioned subset of 39 articles can be viewed at the AsteriX Web site including the uncorrected interpretations (http://swift.cmbi.ru.nl/bitmapb/).

AsteriX was primarily designed to reduce the amount of work when extracting ligand coordinates from pictures from either very old or actually very recent articles. Drawing an intermediate size ligand with a molecular editor such as JME or Cactvs[25] tends to cost a trained person at least a minute, if not much more. Drawing a more complicated structure such as the glycyrrhizic acid shown in Figure 3 took us more than five minutes. Figure 6 shows a series of examples from our test set that were interpreted entirely correct.

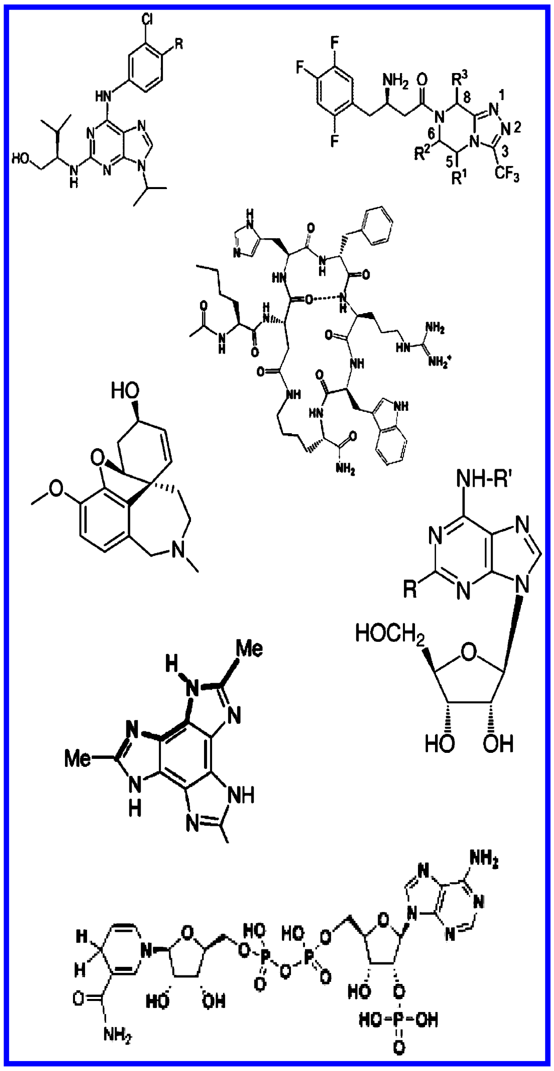
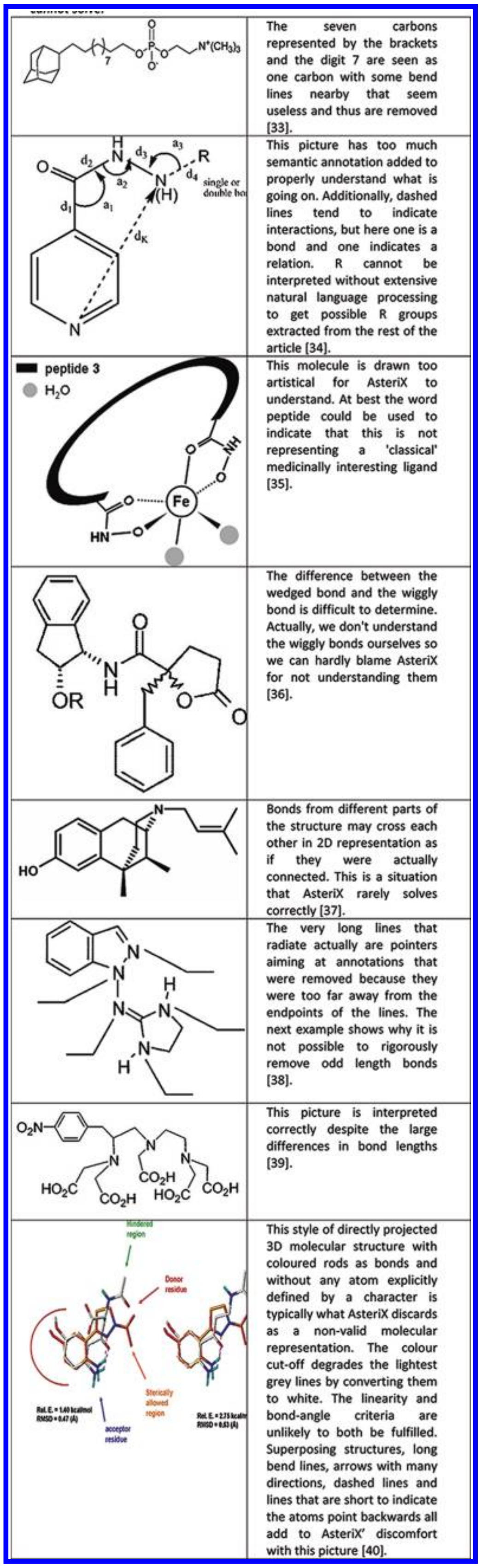


Figure 6. Series of pictures of ligands that were correctly interpreted and converted into 3D coordinates by AsteriX. These structures were extracted (from top to bottom, from left to right) from the refs 26–32.

There is only one way to do things right, but there are a thousand ways to do it wrong. In Figure 7, we tabulate and explain a series of pictures that each for one reason or another could not be interpreted correctly. We have gathered at the AsteriX Web site several more examples of highly complicated representations that illustrate the problems our software cannot resolve yet and, more generally, the challenges that ligand recognition software face.

We have omitted dozens of examples where AsteriX either miraculously manages to solve problems or miserably fails to achieve its goals. All these problems can be amply appraised by visiting the test case section of the Web server at the AsteriX portal http://swift.cmbi.ru.nl/bitmapb/. In the description of AsteriX' algorithm we have, for sake of brevity and readability, left out the discussion of many modules dealing with nitty-gritty



N⁺(CH₃)₃ 7	The seven carbons represented by the brackets and the digit 7 are seen as one carbon with some bend lines nearby that seem useless and thus are removed [33].
single or double bond R (H)	This picture has too much semantic annotation added to properly understand what is going on. Additionally, dashed lines tend to indicate interactions, but here one is a bond and one indicates a relation. R cannot be interpreted without extensive natural language processing to get possible R groups extracted from the rest of the article [34].
	This molecule is drawn too artistical for AsteriX to understand. At best the word peptide could be used to indicate that this is not representing a 'classical' medicinally interesting ligand [35].
	The difference between the wedged bond and the wiggly bond is difficult to determine. Actually, we don't understand the wiggly bonds ourselves so we can hardly blame AsteriX for not understanding them [36].
	Bonds from different parts of the structure may cross each other in 2D representation as if they were actually connected. This is a situation that AsteriX rarely solves correctly [37].
	The very long lines that radiate actually are pointers aiming at annotations that were removed because they were too far away from the endpoints of the lines. The next example shows why it is not possible to rigorously remove odd length bonds [38].
	This picture is interpreted correctly despite the large differences in bond lengths [39].
Hindered region Donor residue Sterically allowed region acceptor residue	This style of directly projected 3D molecular structure with coloured rods as bonds and without any atom explicitly defined by a character is typically what AsteriX discards as a non-valid molecular representation. The colour cut-off degrades the lightest grey lines by converting them to white. The linearity and bond-angle criteria are unlikely to both be fulfilled. Superposing structures, long bend lines, arrows with many directions, dashed lines and lines that are short to indicate the atoms point backwards all add to AsteriX' discomfort with this picture [40].

Figure 7. Examples of problems that AsteriX encountered and often could not solve.

details and parameters that were empirically adjusted. Several of these are listed with some detail at the AsteriX portal in the theory section. The software can freely be used, but we request that users make sensible use of the facility as the interpretation of one PDF can cost up to 60 min CPU time or occasionally even more, especially when articles contain high resolution pictures.

Academic users can obtain AsteriX with source code included free of costs. The AsteriX distribution conditions are available at the AsteriX portal.

■ DISCUSSION

AsteriX was designed to aid with obtaining coordinates of interesting molecular entities observed in literature that are not obtainable from publicly accessible sources. Obviously, the scientific community should strive to achieve that the chemical structure of all small molecules reported in the literature are also deposited electronically in a worldwide database. As it might take a while for this dream to be realized, software like OSRA, CLiDE, and AsteriX will be needed in everyday drug design research. OSRA, CLiDE, and AsteriX are comparable in terms of average Tanimoto index obtained for test sets or in terms of numbers of utterly failed reconstructions. However, they often fail differently for different molecules, so a future combination of these packages in a meta-reconstruction server might lead to a better overall result than each of them can achieve individually. The AsteriX software is available for such purposes.

AsteriX differs from CLiDE and OSRA in that it was not originally designed to interpret patent spreadsheets but has only aimed at extracting ligand data from the literature. AsteriX therefore is integrated in an interactive workbench that allows the user to add his or her personal intelligence to the reconstruction process.

The whole process consists of a series of distinct steps that all are amenable to improvements. Recognizing which pictures contain ligands and which ones do not will probably benefit from a very large scale pattern recognition (neural network or other) approach. This will require a (human) feedback mechanism to generate a very large and guaranteed correct training set.

AsteriX' rather simple character recognition algorithm works surprisingly well, but we do expect that neural network based approaches will be able to improve this step. A better recognition of characters will avoid some of the errors of the type illustrated in Figure 5. Recognizing that something is not a character is equally important, of course, as most pixels can only be used once.

Correctly reconstructing a ligand structure after it has been decomposed into vectors, letters, and numbers turned out to be the most difficult part of the AsteriX project and certainly constitutes its weakest aspects. We have encountered a whole array of highly challenging difficulties, only very few of which could be attempted to be solved. A major problem is to distinguish the elements of the drawing pertaining to the structural representation of the ligand such as bonds, atom labels, and chemical groups from those conveying information relative to structure generalization (substitution pointers, repeat bracket, etc.) or to attached knowledge such as reaction arrows, stereochemical pointers, radicals, or other semantic annotations. A second level of difficulties is the recognition of nonlinear elements such as curved arrows or wiggly bonds. Another major problem arises when bonds cross each other in cage compounds, when parts of compounds are superimposed, or when totally artifactual elements are added to the structure as a form of annotation. Not always do annotations that make the structure clearer to a human also make them clearer to the software.

In evaluating the capabilities of our software, we have attempted to draw a comparison with OSRA, another recently published ligand recognition program. At this moment AsteriX seems to perform slightly better on its own test case and slightly worse on OSRA's published test case, but such comparisons are meaningless as the authors of all these packages keep improving their products and keep coping with the one unpleasant graphical exception after the other.

In the end, it boils down to the same remark that has been in many journals throughout the years: it is always best if the person who generates the data also deposits it in a publicly accessible database. But even when that happens, AsteriX might be useful for all those cases in which the structure is published in a journal before it enters the database.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: v-lounnas@unicancer.fr. Phone: +31 24 3619390 Fax: +31 24 3619395.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Maarten Hekkelman for his precious help in setting up the AsteriX Web server and D. Lutje Hulsik for helping with the software to edit the figures. V.L. thanks the people at the CMBI for their interest and challenging questions as well as their warm hospitality. Stimulating discussions with J. L. Martin and E. Lasonder are acknowledged.

■ REFERENCES

(1) Initative for Chemical Genetics. ChemBank. http://chembank.broad.harvard.edu/ (accessed January 17, 2012).

(2) Chemical Components in PDB. PDB*e*. http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl (accessed January 17, 2012).

(3) http://www.genome.jp/keeg/ligand.html (accessed January 17, 2012).

(4) TimTec Home Page. http://www.timtec.net/ (accessed January 17, 2012).

(5) PubChem Structure Search. PubChem. http://pubchem.ncbi.nlm.nih.gov/search/ (accessed January 17, 2012).

(6) ChEMBL Home Page. http://www.ebi.ac.uk/chembldb/ (accessed January 17, 2012).

(7) JME Molecular Editor Home Page. http://www.molinspiration.com/jme/ (accessed January 17, 2012).

(8) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581. http://www.netsci.org/Resources/Software/Modeling/Conf/corina.html (accessed January 17, 2012).

(9) ChemAxon Home Page. http://www.chemaxon.com/ (accessed January 17, 2012).

(10) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A.; Simpson, R.; Tonnelier, C.; Venczel, T.; Johnson, A. Chemical Literature Data Extractions. The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 338−344.

(11) Valko, A. T.; Johnson, A. P. CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Inf. Model.* **2009**, *49*, 780−787.

(12) Filippov, I. V.; Nicklaus, M. C. Optical structure recognition software to recover chemical information: OSRA, An open source solution. *J. Chem. Inf. Model.* **2009**, *49*, 740−743.

(13) Xpdf-3.00-linux, version 3.00. http://www.foolabs.com/xpdf/ (accessed January 17, 2012).

(14) Lombardo, L. J.; Francis, Y. L.; Chen, P.; Norris, D.; Barrish, J. C.; Behnia, K.; Castaneda, S.; Cornelius, L. A. M.; Das, J.; Doweyko, A. M.; Fairchild, C.; Hunt, J. T.; Inigo, I; Johnston, K.; Kamath, A.; Kan, D.; Klei, H.; Marathe, P.; Pang, S.; Peterson, R.; Pitt, S.; Schieven, G. L.; Schmidt, R. J.; Tokarski, J.; Wen, M. L.; Wityak, J.; Borzilleri, R. M. Discovery of *N*-(2-chloro-6-methyl-phenyl)-2-(6-(4-(2-hydroxyethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), a dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays. *J. Med. Chem.* **2004**, *47*, 6658−6661.

(15) Reichwald, C.; Shimony, O.; Dunkel, U.; Sacerdoti-Sierra, N.; Jaffe, L. C.; Kunick, C. 2-(3-Aryl-3-oxopropen-1-yl)-9-*tert*-butyl-paullones: A new antileishmanial chemotype. *J. Med. Chem.* **2008**, *51*, 659−676.

(16) Hu, H. Y.; Horton, J. K.; Gryk, M. R.; Prasad, R.; Naron, J. M.; Sun, D. A.; Hecht, S. M.; Wilson, S. H.; Mullen, G. P. Identification of small molecule synthetic inhibitors of DNA S polymerase by NMR chemical shift mapping. *J. Biol. Chem.* **2004**, *279*, 39736−39744.

(17) Zimmermann, M.; Thi, L.; Hofmann, M. Combating illiteracy in chemistry: Towards computer-based chemical structure reconstruction. *ERCIM News* **2005**, *60*, 40−41.

(18) Kral, P. Chemical Structure Recognition via an Expert System Guided Graph Exploration. Diploma Thesis in Bioinformatics, Fraunhofer Institute − Ludwig Maximilian Universität/Technische Universität München, 2007.

(19) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(20) PDB File Format. Protein Data Bank (PDB). http://www.rcsb.org/pdb/static.do?p=file_formats/pdb/index.html (accessed January 17, 2012).

(21) SYBYL MOL2 Format. http://www.csb.yale.edu/userguides/datamanip/dock/DOCK_4.0.1/html/Manual.41.html (accessed January 17, 2012).

(22) Joossens, J.; Ali, O. M.; El-Sayed, I.; Surpateanu, G.; Van der Veken, P.; Lambeir, A. M.; Setyono-Han, B.; Foekens, A. J.; Schneider, A.; Schmalix, W.; Haemers, A.; Augustyns, K. Small, potent, and selective diaryl phosphonate inhibitors for urokinase-type plasminogen activator with in vivo antimetastatic properties. *J. Med. Chem.* **2007**, *50*, 6638−6646.

(23) *Journal of Medicinal Chemistry*. American Chemical Society. http://pubs.acs.org/journal/jmcmar (accessed January 17, 2012).

(24) Bender, A.; Mussa, 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.

(25) CACTVS System Home Page. http://www2.chemie.uni-erlangen.de/software/cactvs/index.html (accessed January 17, 2012).

(26) Naumann, T.; Matter, H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes. *J. Med. Chem.* **2002**, *45*, 2366−2378.

(27) Dooseop, K.; Kowalchick, J. E.; Brockunier, L. L.; Parmee, E. R.; Eiermann, G. J.; Fisher, M.; He, H.; Leiting, B.; Lyons, K.; Scapin, G.; Patel, S. B.; Petrov, A.; Pryor, K. D.; Roy, R. S.; Wu, J. K.; Zhang, X.; Wyvratt, M. J.; Zhang, B. B.; Zhu, L.; Thornberry, N. A.; Weber, A. E. Discovery of potent and selective dipeptidyl peptidase IV inhibitors derived from aminoamides bearing subsituted triazolopiperazines. *J. Med. Chem.* **2008**, *51*, 589−602.

(28) Mayorov, A. V.; Cai, M.; Palmer, E. S.; Dedek, M. M.; Cain, J. P.; Van Scoy, A. R.; Tan, B.; Vagner, J.; Trivedi, D.; Hruby, V. J. Structure−activity relationships of cyclic lactam analogues of r-melanocyte-stimulating hormone (r-MSH) targeting the human melanocortin-3 receptor. *J. Med. Chem.* **2008**, *51*, 187−195.

(29) Pietsch, M.; Gutschow, M. Synthesis of tricyclic 1,3-oxazin-4-ones and kinetic analysis of cholesterol esterase and acetylcholinesterase inhibition. *J. Med. Chem.* **2005**, *48*, 8270−8288.

(30) Jacobson, K. A.; Gao, Z. G.; Tchilibon, S.; Duong, H. T.; Joshi, B. V.; Sonin, D.; Liang, B. T. Semirational Design of (north)-methanocarba nucleosides as dual acting A1 and A3 adenosine receptor agonists: Novel prototypes for cardioprotection. *J. Med. Chem.* **2005**, *48*, 8103−8107.

(31) Novellino, E.; Cosimelli, B.; Ehlardo, M.; Greco, G.; Iadanza, M.; Lavecchia, A.; Rimoli, M. G.; Sala, A.; Da Settimo, A.; Primofiore, G.; Da Settimo, F.; Taliani, S.; La Motta, C.; Klotz, K. N.; Tuscano, D.; Trincavelli, M. L.; Martini, C. 2-(Benzimidazol-2-yl)quinoxalines: A novel class of selective antagonists at human A1 and A3 adenosine receptors designed by 3D database searching. *J. Med. Chem.* **2005**, *48*, 8253−8260.

(32) Poirier, D.; Boivin, R. P.; Tremblay, M. R.; Berube, M.; Qiu, W.; Lin, S. X. Estradiol−adenosine hybrid compounds designed to inhibit type 1 17-hydroxysteroid dehydrogenase. *J. Med. Chem.* **2005**, *48*, 8134−8147.

(33) Calogeropoulou, T.; Angelou, P.; Detsi, A.; Fragiadaki, I.; Scoulica, E. Design and synthesis of potent antileishmanial cycloalkylidene-substituted ether phospholipid derivatives. *J. Med. Chem.* **2008**, *51*, 897−908.

(34) Ventura, C.; Martins, F. Application of quantitative structure-activity relationships to the modeling of antitubercular compounds. 1. The hydrazide family. *J. Med. Chem.* **2008**, *51*, 612−624.

(35) Blat, D.; Weiner, L.; Youdim, M. D. H.; Fridkin, M. A Novel iron-chelating derivative of the neuroprotective peptide NAPVSIPQ shows superior antioxidant and antineurodegenerative capabilities. *J. Med. Chem.* **2008**, *51*, 126−134.

(36) Wu, X.; Öhrngren, P.; Ekegren, J. K.; Unge, J.; Unge, T.; Wallberg, H.; Samuelsson, B.; Hallberg, A.; Larhed, M. Two-carbon-elongated HIV-1 protease inhibitors with a tertiary-alcohol-containing transition-state mimic. *J. Med. Chem.* **2008**, *51*, 1053−1057.

(37) Berardi, F.; Ferorelli, S.; Abate, C.; Pedone, M. P.; Colabufo, N. A.; Contino, M.; Perrone, R. Methyl substitution on the piperidine ring of N-[ω-(6-methoxynaphthalen-1-yl)alkyl] derivatives as a probe for selective binding and activity at the σ1 receptor. *J. Med. Chem.* **2005**, *48*, 8237−8244.

(38) Saczewski, F.; Kornicka, A.; Rybczynska, A.; Hudson, A. L.; Miao, S. S.; Gdaniec, M.; Boblewski, K.; Lehmann, A. 1-[(Imidazolidin-2-yl)imino]indazole. Highly r2/I1 selective agonist: Synthesis, X-ray structure, and biological activity. *J. Med. Chem.* **2008**, *51*, 3599−3608.

(39) Chong, H. S.; Ma, X.; Le, T.; Kwamena, B.; Milenic, D. E.; Brady, E. D.; Song, H. A.; Brechbiel, M. W. Rational design and generation of a bimodal bifunctional ligand for antibody-targeted radiation cancer therapy. *J. Med. Chem.* **2008**, *51*, 118−125.

(40) Entrena, A.; Camacho, M. E.; Carrion, M. D.; Lopez-Cara, L. C.; Velasco, G.; Leon, J.; Escames, G.; Acuna-Castroviejo, D.; Tapias, V.; Gallo, M. A.; Vivo, A.; Espinosa, A. Kynurenamines as neural nitric oxide synthase inhibitors. *J. Med. Chem.* **2005**, *48*, 8174−8181.