JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

# NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats

Sascha Urbaczek,[†] Adrian Kolodzik,[†] J. Robert Fischer,[†] Tobias Lippert,[†] Stefan Heuser,[‡,∥] Inken Groth,[‡] Tanja Schulz-Gasch,[§] and Matthias Rarey*[,†]

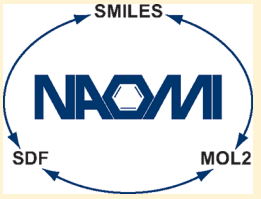
[†]Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Research Active Ingredients, Beiersdorf AG, Troplowitzstrasse 15, 22529 Hamburg, Germany

[§]Pharmaceutical Division, F. Hoffmann-La Roche Ltd., CH-4070 Basel, Switzerland

Supporting Information

**ABSTRACT:** In most cheminformatics workflows, chemical information is stored in files which provide the necessary data for subsequent calculations. The correct interpretation of the file formats is an important prerequisite to obtain meaningful results. Consistent reading of molecules from files, however, is not an easy task. Each file format implicitly represents an underlying chemical model, which has to be taken into consideration when the input data is processed. Additionally, many data sources contain invalid molecules. These have to be identified and either corrected or discarded. We present the chemical file format converter NAOMI, which provides efficient procedures for reliable handling of molecules from the common chemical file formats SDF,[1] MOL2,[2] and SMILES.[3] These procedures are based on a consistent chemical model which has been designed for the appropriate representation of molecules relevant in the context of drug discovery. NAOMI's functionality is tested by round robin file IO exercises with public data sets, which we believe should become a standard test for every cheminformatics tool.




## INTRODUCTION

Chemical file formats provide the necessary data for application programs and offer a means to share results with other scientists in a computer readable form. For small molecules, the most commonly used formats are Symyx SDF V2000 (formerly MDL SDF),[1] Tripos MOL2,[2] and Daylight SMILES.[3] Virtually all public databases provide molecular files of at least one of these types.

Unfortunately, many programs do not accept all formats as input or generate only some of them as output. Hence, file format converters are needed to exchange data between these tools. This becomes especially important if several of these tools are combined in a workflow. The consistent conversion of molecules is crucial at this point, since even minor alterations might result in errors in subsequent calculations.

The conversion process is difficult and error prone. File formats implicitly represent an underlying chemical model which has to be taken into account. Hence, the file format conversion is actually a conversion between different chemical representations. Furthermore, some programs generate files that do not follow format specifications or contain errors. Converters must thus be able to identify errors and ambiguities in input data and resolve them consistently or discard the corresponding molecule.

Since chemical file formats play such a central role in cheminformatics, every tool and software package must be able to read and write molecular files. Hence, every tool that supports more than one file format can be used as a converter. However, there are tools which have specifically been designed for file format conversion, such as the free software OpenBabel[4] and, more recently, fconv[5] or the commercial tools MOL2Mol,[6] MN. Convert,[7] and Babel.[8] Furthermore, there is a large number of programming libraries for cheminformatics, both open source and proprietary, which provide the necessary functionality to read and write molecules. Evidently, these can be used to implement converter tools. Examples of such libraries are Open Babel,[9] CDK,[10] CACTVS,[11] JOELib,[12] PerlMol,[13] OEChem,[14] and RDKit.[15] Additionally, some tools are routinely used for file format conversions, although that is not their specific purpose. Typical examples are programs for the generation of 3D coordinates, such as CORINA,[16] LigPrep,[17] and CONCORD.[18]

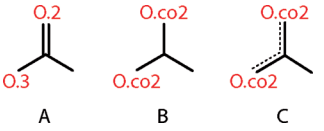



**Figure 1.** Different representations of carboxylates as observed in MOL2 files.

We have implemented a new tool for the consistent conversion of chemical file formats called NAOMI. This converter is based on a robust chemical model which is designed to appropriately describe organic molecules relevant in the context of drug discovery. It provides a reliable and accurate internal representation which allows for a consistent interconversion of the widely used

**Received:** July 14, 2011
**Published:** November 08, 2011

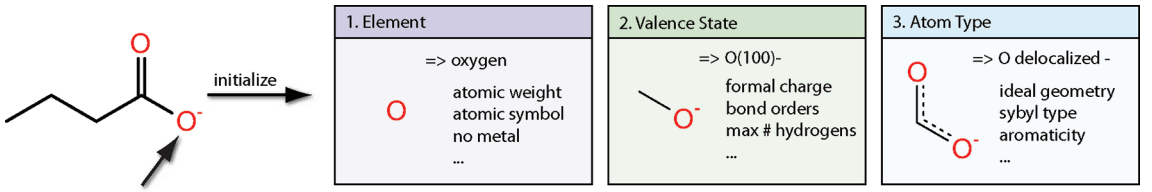



**Figure 2.** Annotation of the three levels of chemical information for an oxygen of a carboxylate.

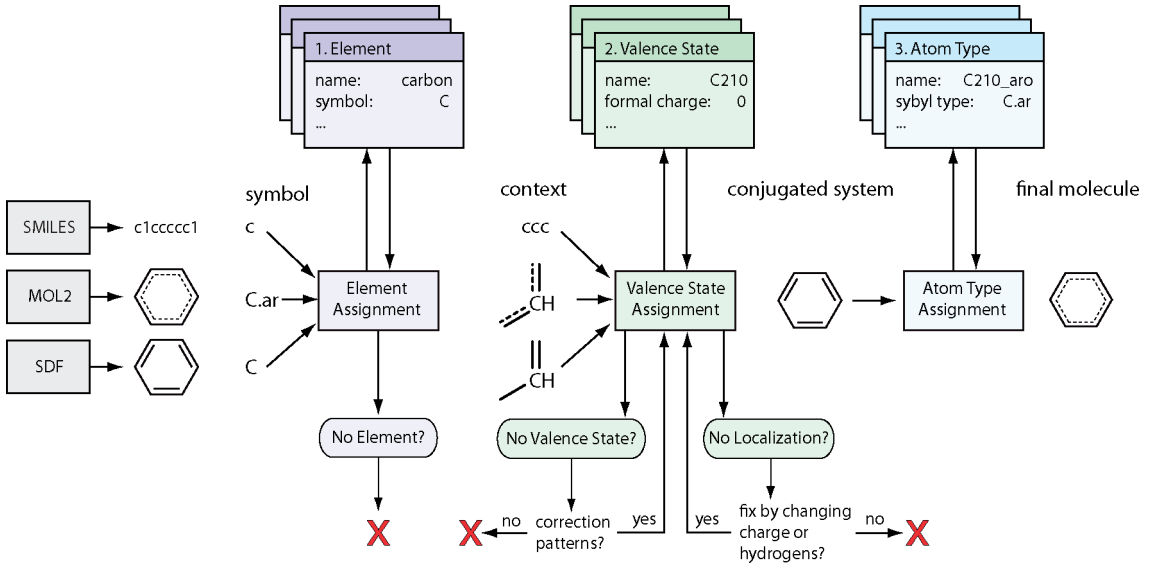



**Figure 3.** Schematical view of the three steps of molecule initialization.

molecular file formats SDF V2000,[1] MOL2,[2] and SMILES.[3] NAOMI also supports reading and writing SDF V3000 files but does currently not implement all associated features, e.g., self-contained sequence representation. NAOMI checks the chemical validity of molecules and calculates molecular descriptors independent of input file formats.

Although file IO is a task all cheminformatics tools have to perform, not very much is known about the methodologies applied to address the problems related to file conversion. We assume that many tools use approaches very similar to NAOMI, but unfortunately these are mostly not published. Furthermore, file IO and conversion is rarely tested and validated exhaustively. The aim of this paper is to explicitly put the focus on these tasks to demonstrate the complexity and typical pitfalls. We present a round robin test for cheminformatics tools able to read and write different file formats and advocate the use of such tests routinely.

**File Format Conversion.** The conversion of file formats involves two steps: First, the information provided by the input format is interpreted to build an internal representation of the molecule. Second, all relevant data for the target format is derived from this representation. Due to the different underlying chemical models of the file formats, the conversion usually involves switching from one chemical description to another. Thus, it is important to consider the requirements and limitations of these descriptions.

The Symyx SDF format[1] represents molecules by a single valence bond structure, also called Lewis structure.[19] Hydrogens are frequently omitted to save disk space, while the file format specification ensures the presence of formal charges. The valence bond description has limitations concerning kekule and resonance

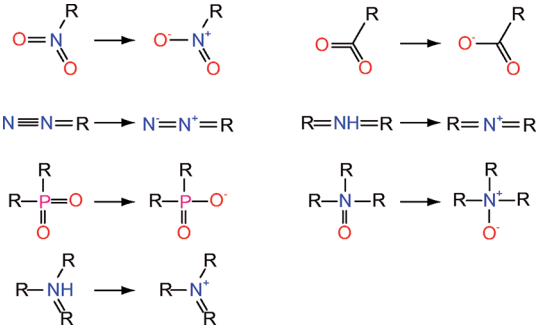


**Figure 4.** If no valence state can be identified for an atom, then a set of simple correction patterns is applied.

structures, since multiple equivalent valence bond forms of the same molecule may exist.

SMILES[20] can represent molecules by a single valence bond structure, whereas hydrogens are virtually always omitted. The format also implements the concept of aromatic atoms and bonds, which allows to represent aromatic systems with different equivalent kekule forms by a single delocalized description. According to the Daylight theory manual,[20] aromaticity in SMILES is however not intended to model physicochemical properties (Daylight theory manual, page 14). Nevertheless, aromatic atoms and bonds are commonly used to describe molecules which are aromatic in a chemical sense, although a single valence bond structure would be sufficient to characterize these molecules unambiguously.

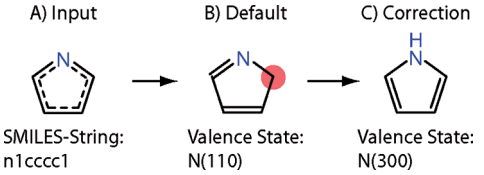



**Figure 5.** If an input file annotates aromatic atoms and bonds (A), default valence states are assigned in a first step (B). If this attempt is not successful, alternative valence states are considered (C) to correct the input.

The TRIPOS MOL2 format implements the concept of aromatic atoms and bonds, too. Furthermore, the format offers the possibility to describe equivalent resonance forms of common functional groups, such as carboxylates and guanidinium groups, with a delocalized representation. This is realized using specific atom types, called sybyl types, which include information about the atom's hybridization. Usually, MOL2 files do not provide formal charges, but hydrogens are specified. Unfortunately, there is no exact documentation on how the sybyl types must be assigned. This leads to considerable differences between MOL2 files written by different tools. As shown in Figure 1 there are many ways to combine sybyl types, bond orders, and charges to describe the same functional group.

## METHODOLOGY

**Chemical Model.** A consistent chemical model is the keystone for an appropriate internal representation of molecules in cheminformatics application. It also provides the framework for the identification and correction of erroneous input molecules.

The atom-centered chemical model of NAOMI comprises three different levels of chemical information which are assigned to each atom during an initialization procedure. Each level extends the environment that is considered and provides a more detailed description of the atom.

The element is the first and most basic level of description. It provides properties which depend only on the atom's chemical element. These properties comprise the element symbol, the element name, the atomic number, the atomic weight, the van der Waals radius, the number of valence electrons, the covalent radius, and whether the element is considered a metal.

The valence state is the second level of chemical information and extends the scope of the chemical element by taking bonds and formal charges into account. Each valence state represents a valid bond pattern of an atom in a valence bond structure of the molecule. Valence states contain topological information which include formal charge, number of bonds, bond orders, number of free electrons, and whether the corresponding atom can be part of a conjugated or aromatic system.

The atom type extends the valence state to model effects, such as aromaticity and the existence of equivalent resonance forms. This is needed to compensate for the shortcomings of a localized molecular description.

To determine an atom type, the atom and all atoms in its conjugated system (if applicable) are considered. Atom types provide an ideal geometry, a corresponding sybyl type, mark atoms as conjugated or aromatic, and contain information about delocalized electrons. Additionally, an atom type marks the corresponding atom as a hydrogen-bond acceptor or as a potential hydrogen-bond donor.

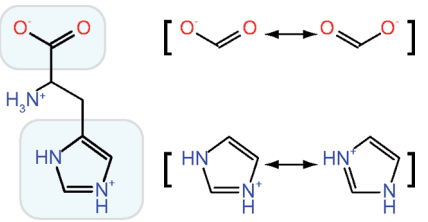


**Figure 6.** Molecules are partitioned into zones of conjugated atoms. The two oxygen atoms of the carboxylate group and the two nitrogen atoms of the imidazole ring have different valence states but identical atom types. Therefore, the valence states describe a localized structure with a defined formal charge, and the atom types describe a delocalized structure, with a delocalized charge.

Each atom is assigned a corresponding element, valence state and atom type (see Figure 2). Valence states ensure that each molecule has a valid valence bond structure, while atom types allow easy access to a delocalized description.

The basic assumption of the chemical model is that organic molecules which are relevant in the drug discovery context can always be represented by at least one valence bond structure. If that is not the case, then the molecule will either be corrected or discarded. Since there are no strict valence rules for metallic elements, only monatomic ions are accepted. Molecules containing covalently bound metals are currently not supported by the model.

**Molecule Initialization.** *Overview.* During the molecule initialization data from input files is used to build the internal representation of the molecule. This task is carried out in three separate steps (see Figure 3).

*Element Assignment.* First, the molecular graph is built from the connectivity data provided by the input file. During this process, the element for each atom is determined, and initial bond types are assigned. The perception of elements, bond types, and connectivity from the different file formats is implemented according to their respective specifications. All elements of the periodic table and bonds of type single, double, triple, and aromatic are supported. Molecules which have atoms or bonds with undefined types are discarded at this point, since this information is required in the subsequent steps.

The initial data are used to generate a valid valence bond form of the molecule. A valence bond form is valid if valence states can be assigned to all atoms and the aromatic bonds can be localized. If no valence bond form can be generated and no correction is possible, the molecule is discarded.

*Valence State Assignment.* They are selected on basis of the formal charge and bond orders of the atom. Hence, molecules with formal charges, hydrogens, and localized bond orders are the optimal input for this procedure. In this case, the assignment is straightforward and unambiguous. The omission of hydrogens or the use of aromatic bonds, which basically corresponds to the omission of bond orders, also poses no problem, since the remaining properties are still sufficient to reach an unambiguous assignment. If charges or multiple properties are missing, then additional data from the input format is necessary to resolve ambiguities.

Each file format makes use of a different molecular representation and applies certain strategies to omit redundant information. Hence, individual assignment procedures are needed for each file format.

Molecules from SDF are supplied in a valence bond form, which allows a direct comparison to valence states. If hydrogens are

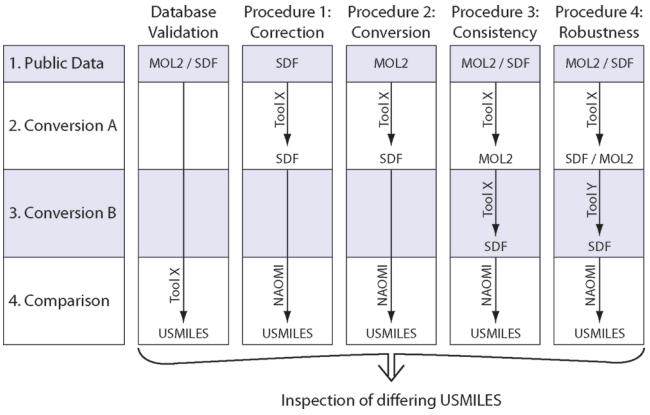



**Figure 7.** Various procedures test different aspects of file format conversions. During these procedures, molecules are converted by different combinations of tools. USMILES are used for the comparison of the resulting molecules.

**Table 1. Options Used for Computing Time Benchmarks**

| tool/options | explanation |
|---|---|
| CORINA | |
| −d wh | write hydrogens to output file |
| −d no 3d | disable generation of 3D coordinates |
| −t n | do not write trace file |
| MOE | |
| −SVL script | (see Supporting Information) |
| NAOMI | |
| −v 0 | do not print messages to shell |
| Open Babel | |
| −o can | generate USMILES (only for SMILES as output) |

omitted, formal charges and multiple bonds are sufficient to unambiguously identify the correct valence states.

Molecules from SMILES may provide information on the bond orders explicitly, whereas hydrogens are virtually always omitted. If this is the case, the assignment works the same way as for SDF. Additionally, SMILES implements the concept of aromatic bonds. This means that bond orders and hydrogens can be missing, and hence ambiguities arise for certain types of atoms. The most prominent example is the pyrrole-like aromatic nitrogen (see Figure 5) which has to be provided with explicit hydrogens for an unambiguous assignment.

Molecules from MOL2 usually have all hydrogens attached but lack the specification of formal charges. If they also contain aromatic bonds, two properties are missing. These ambiguities can only be resolved by using sybyl type information. Additionally, some resonance forms of common functional groups are indicated by specific sybyl types. Their bond types and valence states are adapted accordingly in a postprocessing step.

If no valence state could be found for an atom, the atom's environment is checked by using simple patterns representing common valence errors (see Figure 4). If a pattern matches, then a valence state is assigned, and the bond orders and valence states of the environment are adapted. Otherwise the molecule is discarded.

Afterward, the bonds marked as aromatic in the input file are localized to ensure a valid valence bond form. Information about the localized bond orders for each atom is provided by its corresponding valence state. The information is used in a recursive algorithm to assign defined bond orders to all bonds.

**Table 2. Validation of Input Data Sets by NAOMI**

| data set | no. molecules | no. rejected molecules | corrected molecules | no. diffs MOL2 ↔ SDF |
|---|---|---|---|---|
| DUD ligands[23] | 3961 | 0 | 10 | 0 |
| DUD decoys[23] | 124 413 | 1 | 13 | 0 |

If the assignment of bond orders was not successful using the default valence states, all atoms of a molecule are checked for an alternative valence state assignment using rule sets specific to the respective file formats (see Figure 5). All combinations of these alternatives are enumerated, and the most probable solution is picked by a simple scoring scheme. The score is calculated as the sum of atoms which have the same valence states with respect to the initial structure. Thus, the procedure assures a minimum deviation from the default assignment. If there are multiple solutions with equal scores, a canonical solution is picked. If no solution could be found, then the molecule is discarded.

*Atom Type Assignment.* At this point, a valid valence bond form of the molecule is available and can be accessed during subsequent calculations. Since all necessary information can now be derived from the internal representation, the following steps are independent of the input file format.

The next step is the generation of a delocalized description for the molecule. The description allows to overcome the limitations of the valence bond representation concerning kekule and resonance structures. Although these aspects are handled by separate procedures, both need information about the molecule's rings. These are calculated using the relevant cycles algorithm as described by Vismara.[21]

Since equivalent kekule structures can only occur in cyclic systems, this information is stored directly in the molecule's rings. A ring is marked as delocalized if it has alternating single and double bonds and the number of delocalized electrons does fulfill Hueckel's rule. Bonds from rings which are already marked are considered both single and double during the check of neighboring rings. To ensure that the assignment for all rings is independent from the initial valence bond form, the assignment procedure is repeated until the total number of marked rings does not change anymore.

For the identification of equivalent resonance forms, the molecule is partitioned into zones which correspond to its conjugated systems (see Figure 6). This is done by using the information provided by the valence states in combination with the molecule's rings. Each zone is checked for pairs of atoms for which a formal charge can be exchanged. These atoms can be identified by comparison of their corresponding valence states. Then all possible resonance forms are enumerated, and all atoms with delocalized charges are marked. Finally, suitable atom types are selected from a list provided by the valence state using the information about the conjugated system and the delocalization of the atom.

After the initialization procedure, the molecule is represented by a valence bond description (valence states and bond orders) and a delocalized description (atom types and delocalization flags). Both descriptions can be used in subsequent steps.

**Validation.** To evaluate the quality of file format conversions, a method for comparing input and converted molecules is required. Unfortunately, there is no direct way to determine if two molecular representations are identical. This is especially true if they are stored in different file formats.

The comparison of unique SMILES (USMILES)[22] is an easy and verifiable way to identify differences between molecules. Two things have to be taken into consideration with this approach: First, USMILES generated with different tools are often not identical. This means that the method will only be reliable if the USMILES come from the same source. Second, some file format specific information will be lost during the conversion. Therefore, USMILES should be obtained from SDF files, since it provides an unambiguous valence bond structure.

The public DUD ligand and the DUD decoy[23] data sets are used in all validation procedures. To establish a reference for the comparison, both were converted from SDF and MOL2 to USMILES (see Figure 7). These USMILES serve as a basis to determine whether molecules change during conversion steps.

To investigate a tool's ability to convert file formats, four validation procedures are used as shown in Figure 7. In the first procedure, the internal error correction of the tools is analyzed by conversion of molecules from SDF to SDF. The ability to convert molecules from one format into another is investigated in the second procedure by converting molecules from MOL2 to SDF. The third procedure focuses on a tool's internal consistency by converting back and forth using the same tool twice. Finally, the robustness is checked by using different tools subsequently in a pipeline.

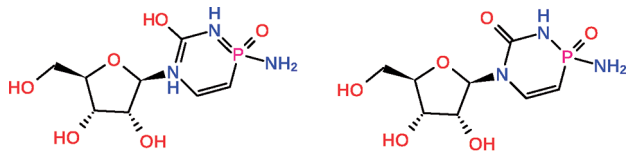



**Figure 8.** Molecule ZINC0153034: (A) Rejected by NAOMI in DUD decoy data set and (B) in current ZINC database.

**Table 3. Data Sets Converted To USMILES by MOE and Open Babel[a]**

| | | | MOL2 ↔ SDF | |
|---|---|---|---|---|
| tool | data set | no. rejected molecules | no. diffs | % of data |
| MOE | DUD ligands | 0 | 1598 | 40% |
| | DUD decoys | 0 | 67 042 | 54% |
| Open Babel | DUD ligands | 0 | 1875 | 47% |
| | DUD decoys | 0 | 46 987 | 38% |

[a] Shown are the differences between the generated USMILES originating from MOL2 and SDF.

**Table 4. Investigation of Correction Functionality**

| | DUD decoys | | DUD ligands | |
|---|---|---|---|---|
| tool | no. rejected | no. corrected | no. rejected | no. corrected |
| CORINA | 0 | 0 | 0 | 0 |
| MOE | 0 | 13 | 0 | 8 |
| NAOMI | 1 | 13 | 0 | 10 |
| Open Babel | 0 | 0 | 0 | 0 |

All validation procedures are performed with CORINA,[24] MOE,[25] Open Babel,[4] and NAOMI. CORINA is commonly used for generating 3D coordinates and for molecular file format conversion and is considered the gold standard. MOE is used for a variety of applications in drug design and supports preparation of ligands for subsequent calculations. This includes the generation of protonation states and tautomers as well as filtering according to molecular descriptors. Correct and consistent reading and writing of molecules forms the basis for these applications. An open source alternative to these tools is Open Babel. Open Babel supports a variety of molecular file formats and is designed to be used as a file format converter.

**Computing Time Benchmarks.** Although the consistency and the quality of the converted molecules are of superior importance, computing times play a significant role due to the increasing sizes of current data sets. Hence, the runtime behavior is analyzed in order to assess their applicability in large setups.

To investigate NAOMI's performance, the ZINC-everything data set is converted from and to MOL2, SDF, and USMILES. Measured computing times are compared to the commonly used tools CORINA, Open Babel, and MOE. For an unbiased comparison, optional settings of these tools are selected to yield similar results compared to NAOMI. Therefore, generation of USMILES and writing of hydrogens are enforced, and output of additional information is minimized (see Table 1). Conversion from SMILES to MOL2 and SDF using CORINA is omitted since CORINA automatically generates 3D coordinates upon conversion. Furthermore, SMILES is not supported as an output format by CORINA. Although, NAOMI is able to conduct its calculations in parallel, this option is disabled for an easier comparison. All file format conversions are performed on a Linux PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

## RESULTS

**Data Set Validation.** Results of the validation of the DUD ligand and DUD decoy data sets[23] are shown in Table 2. NAOMI successfully converts all molecules except one from MOL2 and SDF to USMILES. A small number of incorrectly protonated nitrogens are corrected. One molecule (ZINC1583034) is rejected, as it contains invalid phosphorus and nitrogen atoms (see Figure 8) which cannot be corrected and localized. Since USMILES

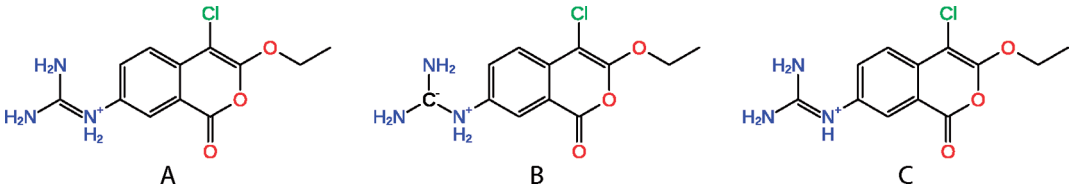



**Figure 9.** (A) Molecule from DUD ligand data set. (B) Corrected molecule from MOE. (C) Corrected molecule from NAOMI.


dx.doi.org/10.1021/ci200324e |J. Chem. Inf. Model. 2011, 51, 3199–3207

generated by NAOMI are identical for both file formats, they can serve as a reference for the following validation procedures.

Both data sets could also be successfully converted to USMILES by MOE and Open Babel. The molecule which was rejected by NAOMI is neither discarded nor corrected by both tools. USMILES originating from MOL2 and SDF, however, differ significantly (see Table 3).

**Table 5. Investigation of Conversion Functionality**

| | DUD decoys | | DUD ligands | |
|---|---|---|---|---|
| tool | no. diffs | % of data | no. diffs | % of data |
| CORINA | 5522 | 4% | 439 | 11% |
| MOE | 4287 | 3% | 181 | 5% |
| NAOMI | 0 | 0% | 0 | 0% |
| Open Babel | 13 469 | 11% | 966 | 24% |

**Table 6. Investigation of tool consistency**

| | | DUD decoys | | DUD ligands | |
|---|---|---|---|---|---|
| tool | starting file format | no. diffs | % of data | no. diffs | % of data |
| CORINA | MOL2 | 5522 | 4% | 439 | 11% |
| | SDF | 4174 | 3% | 235 | 6% |
| MOE | MOL2 | 5770 | 5% | 457 | 12% |
| | SDF | 5683 | 5% | 453 | 11% |
| NAOMI | MOL2 | 0 | 0% | 0 | 0% |
| | SDF | 0 | 0% | 0 | 0% |
| Open Babel | MOL2 | 17 351 | 14% | 1168 | 29% |
| | SDF | 17 364 | 14% | 1168 | 29% |

**Tool Validation 1: Correction.** As mentioned above, the DUD data sets contain 24 invalid molecules in total of which one has been rejected and 23 could be corrected. CORINA and Open Babel convert those without performing any error correction (Table 4). MOE and NAOMI correct the nitrogens with invalid protonation states with differing results (see Figure 9 for an example). Additionally, NAOMI corrects invalid phosphate groups.

**Tool Validation 2: Conversion.** Results of the investigation of the conversion functionality (see Figure 7) are shown in Table 5. By inspection of the differing molecules, we were able to identify a small number of error classes that will be discussed for every tool:

CORINA places positive charges on carbon atoms of guanidinium- and amidinium-like groups. This error also occurs in five-membered aromatic rings containing this substructure.

MOE places positive charges on carbon atoms of guanidinium- and amidinium-like groups in five-membered aromatic rings. Depending on the substituents, the carbon atom is either charged twice or a carbon atom next to it is negatively charged.

Open Babel's most prominent class of errors is the incorrect conversion of aromatic systems containing charged nitrogen atoms. All bonds in these systems are converted to single bonds in the resulting SDF file. The second kind of error concerns protonation states. Open Babel does not consider input hydrogens to determine formal charges. Therefore, many atoms are neutralized during the conversion process. Since MOL2 entries often do not provide formal charges, this may lead to unexpected results.

**Table 7. Investigation of Tool Robustness**

| | | | DUD decoys | | DUD ligands | |
|---|---|---|---|---|---|---|
| tool X | tool Y | starting file format | no. diffs | % of data | no. diffs | % of data |
| CORINA | MOE | MOL2 | 4265 | 3% | 176 | 4% |
| | | SDF | 5931 | 5% | 449 | 11% |
| | NAOMI | MOL2 | 58 | 0% | 0 | 0% |
| | | SDF | 4149 | 3% | 235 | 6% |
| | Open Babel | MOL2 | 5522 | 4% | 439 | 11% |
| | | SDF | 19 192 | 15% | 1371 | 35% |
| MOE | CORINA | MOL2 | 6755 | 5% | 504 | 13% |
| | | SDF | 4656 | 4% | 245 | 6% |
| | NAOMI | MOL2 | 3159 | 3% | 167 | 4% |
| | | SDF | 4585 | 4% | 239 | 6% |
| | Open Babel | MOL2 | 4483 | 4% | 174 | 4% |
| | | SDF | 19 311 | 16% | 1374 | 35% |
| NAOMI | CORINA | MOL2 | 0 | 0% | 0 | 0% |
| | | SDF | 643 | 1% | 17 | 0% |
| | MOE | MOL2 | 176 | 0% | 0 | 0% |
| | | SDF | 1217 | 1% | 221 | 6% |
| | Open Babel | MOL2 | 0 | 0% | 0 | 0% |
| | | SDF | 14 172 | 11% | 1164 | 29% |
| Open Babel | CORINA | MOL2 | 29 896 | 24% | 1887 | 48% |
| | | SDF | 10 047 | 8% | 289 | 7% |
| | MOE | MOL2 | 13 693 | 11% | 973 | 25% |
| | | SDF | 43 285 | 35% | 1703 | 43% |
| | NAOMI | MOL2 | 13 469 | 11% | 966 | 24% |
| | | SDF | 1790 | 1% | 24 | 1% |

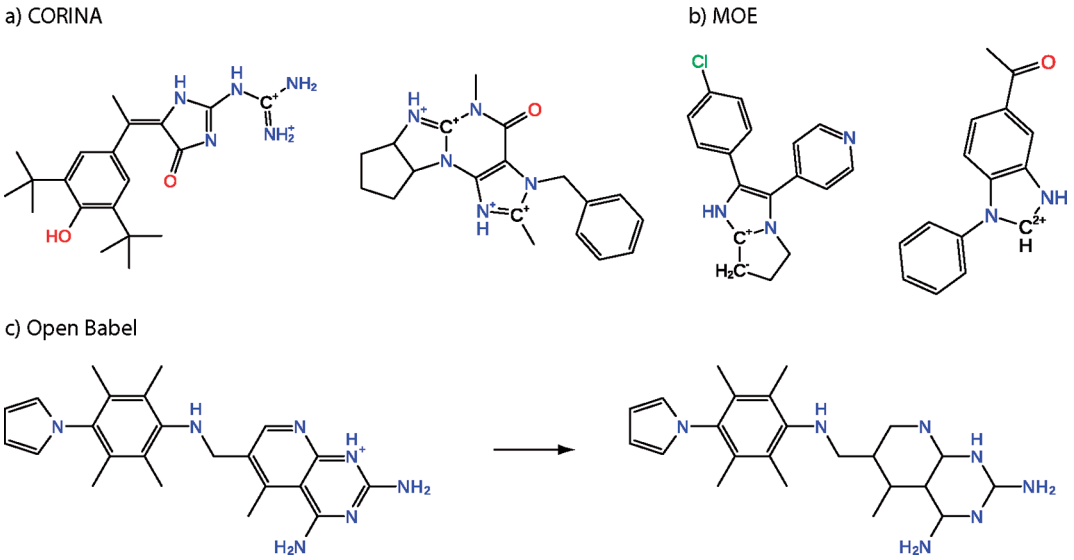



**Figure 10.** Examples of conversion problems with CORINA, MOE, and Open Babel.

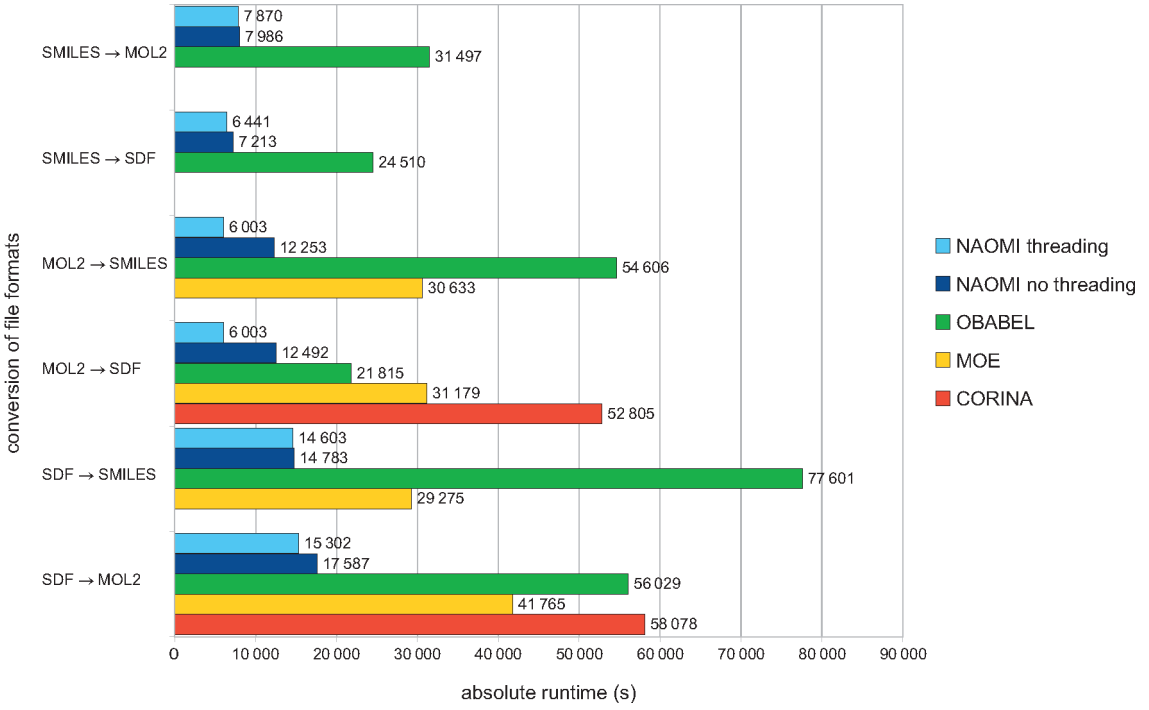



**Figure 11.** Computing times (wall clock time) for file format conversion of the ZINC-everything data set. For CORINA and MOE, only the computation from SDF and MOL2 are comparable, since the conversion from SMILES includes 3D coordinate generation which is not the case for NAOMI and OpenBabel. Furthermore, CORINA does support SMILES as output format.

**Tool Validation 3: Consistency.** Results of the investigation of consistency (see Figure 7) are shown in Table 6. Starting from MOL2, the numbers of differences should be identical to those of validation procedure 2 (see Table 5), since no additional file format conversion is performed. A higher number of errors indicates inconsistencies in reading and writing from and to MOL2. Starting from SDF, no differences at all should occur.

CORINA and NAOMI convert molecules consistently in both cases. The differences which were observed for CORINA when the first input was provided from SDF are introduced by switching from a delocalized to a localized description. Nevertheless, they only represent different valid resonance forms of the original data and are therefore not considered conversion errors. MOE and Open Babel show inconsistencies in both cases.

**Tool Validation 4: Robustness.** The robustness of the investigated tools is analyzed by combining two different tools in a pipeline. Since tools tend to interpret input from file formats differently, the molecules can change with each additional tool included in the workflow. Table 7 indicates that inconsistencies during file format conversion are not uncommon and depend both on the kind of tools used and on the order in which they are combined.


dx.doi.org/10.1021/ci200324e |J. Chem. Inf. Model. 2011, 51, 3199–3207

Furthermore, the success of the conversion strongly depends on the source of the input data. The experiment clearly shows that all tools benefit significantly from preprocessing data sources with NAOMI toward consistency and high quality (see Figure 10).

**Computing Time Benchmarks.** Figure 11 summarizes the computing times for conversion of the ZINC-everything data set. Since NAOMI is designed for large scale cheminformatics applications, it is not surprising that it is substantially faster than the modeling platform MOE. NAOMI supports multithreading resulting in a speed-up by another factor of 1.4. For SDF and SMILES, file IO is usually the rate-determining step. Therefore, threading does not lead to an improvement of runtimes. The MOL2 format however needs a more advanced initialization procedure, thus leading to gains in runtimes when threading is enabled.

In summary, NAOMI achieves a conversion speed of up to 2841 molecules/second on a PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

## CONCLUSION

Handling chemical structures is and remains a complex task. File formats contain chemical descriptions at different levels of detail and are therefore not easy to convert. Since the description of file formats are sometimes ambiguous when it comes to details, software tools tend to interpret them differently. This in turn causes errors in data sets and misinterpretations in tools. For the cheminformatics community, it would be a great benefit to build clear standards for file formats and to certify software with respect to these standards.

Meanwhile, it is important that software tools are at least self-consistent when reading and writing file formats. Evidently, errors in reading molecules from files usually have a substantial impact on downstream algorithms and methods. NAOMI will most certainly have flaws of its own, and in order to find them, constistency checks as those presented are needed. We urge that more of these tests should be published and that the existing ones become a standard validation procedure for all cheminformatics applications.

The command-line converter NAOMI has been implemented in C++ and can be downloaded at http://www.zbh.uni-hamburg.de/naomi. It will be available free of charge for academic use. A convenient graphical user interface for NAOMI's functionality will soon be provided by the chemical library preprocessor MONA (see http://www.zbh.uni-hamburg.de/mona).

## ASSOCIATED CONTENT

**Supporting Information.** Original and corrected structures for both DUD data sets are provided. The corrected structures are supplied in the same file format as the respective input files (SDF or MOL2). Furthermore, a text file containing the SVL commands used for the computations with MOE is supplied. This material is available free of charge via the Internet at http://pubs.acs.org/.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

### Present Address

‖Current address: Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany.

## ACKNOWLEDGMENT

The authors thank Stefan Wefing for his initial ideas concerning the chemical model, Dr. Holger Claußen for testing, Rene Kraus for IO support, and Matthias Hilbig for supplying a graphical interface.

## REFERENCES

(1) *Symyx CTfile Formats*; http://www.symyx.com/downloads/public/ctfile/ctfile.jsp, (accessed January 27, 2011).

(2) *TRIPOS Mol2 File Format*; http://tripos.com/data/support/mol2.pdf, (accessed January 27, 2011).

(3) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(4) *The Open Babel Package*, version 2.3.0; http://openbabel.org, (accessed January 18, 2011),

(5) Neudert, G.; Klebe, G. fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27*, 1021–1022.

(6) *Mol2Mol*; http://www.gunda.hu/mol2mol/index.html, (accessed January 27, 2011).

(7) *MN.Convert*; Molecular Networks GmbH - Computerchemie: Erlangen, Germany; http://www.molecular-networks.com/products/convert, (accessed January 27, 2011)

(8) *Babel*; OpenEye Scientific Software, Inc.: Santa Fe, NM; http://www.eyesopen.com/docs/babel/current/pdf/BABEL.pdf, (accessed January 27, 2011).

(9) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.

(10) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(11) Ihlenfeldt, W.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.

(12) *JOELib/JOELib2*; http://sourceforge.net/projects/joelib/, (accessed January 27, 2011).

(13) *PerlMol*; http://www.perlmol.org/, (accessed January 27, 2011).

(14) *OEChem*; OpenEye Scientific Software, Inc.: Santa Fe, NM; http://www.eyesopen.com/oechem-tk, (accessed January 27, 2011).

(15) *RDKit*; http://rdkit.org/, (accessed Jan 27, 2011).

(16) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(17) *LigPrep*;Schrödinger, LLC: Cambridge, MA; http://www.schrodinger.com/products/14/10/, (accessed January 27, 2011).

(18) *Concord*; Tripos: St. Louis, MO; http://tripos.com/data/SYBYL/Concord_072505.pdf, (accessed January 27, 2011).

(19) Lewis, G. N. The Atom and the Molecule. *J. Am. Chem. Soc.* **1916**, *38*, 762–785.

(20) *Daylight theory manual*, Daylight version 4.9; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA; http://www.daylight.com/dayhtml/doc/theory/index.pdf, (accessed March 8, 2011).

(21) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1–15.

(22) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(23) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801 Data sets (SDF and Mol2) downloaded April 12, 2011.

(24) *CORINA - Fast Generation of High-Quality 3D Molecular Models*, version 3.48; Molecular Networks GmbH - Computerchemie : Erlangen, Germany; http://www.molecular-networks.com/products/corina, (accessed January 18, 2011).

(25) *MOE*, version 2010.10; Chemical Computing Group: Montreal, Quebec, Canada; http://www.chemcomp.com/software.htm, (accessed January 18, 2011).