JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

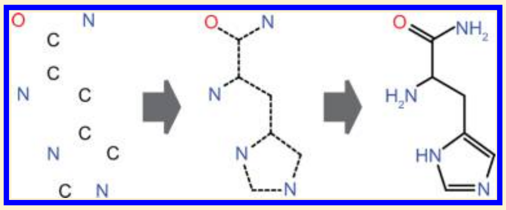
Reading PDB: Perception of Molecules from 3D Atomic Coordinates

Sascha Urbaczek,[†] Adrian Kolodzik,[†,∥] Inken Groth,[‡] Stefan Heuser,[‡,§] and Matthias Rarey*,[†]

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

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

S Supporting Information



ABSTRACT: The analysis of small molecule crystal structures is a common way to gather valuable information for drug development. The necessary structural data is usually provided in specific file formats containing only element identities and three-dimensional atomic coordinates as reliable chemical information. Consequently, the automated perception of molecular structures from atomic coordinates has become a standard task in cheminformatics. The molecules generated by such methods must be both chemically valid and reasonable to provide a reliable basis for subsequent calculations. This can be a difficult task since the provided coordinates may deviate from ideal molecular geometries due to experimental uncertainties or low resolution. Additionally, the quality of the input data often differs significantly thus making it difficult to distinguish between actual structural features and mere geometric distortions. We present a method for the generation of molecular structures from atomic coordinates based on the recently published NAOMI model. By making use of this consistent chemical description, our method is able to generate reliable results even with input data of low quality. Molecules from 363 Protein Data Bank (PDB) entries could be perceived with a success rate of 98%, a result which could not be achieved with previously described methods. The robustness of our approach has been assessed by processing all small molecules from the PDB and comparing them to reference structures. The complete data set can be processed in less than 3 min, thus showing that our approach is suitable for large scale applications.

■ INTRODUCTION

Crystal structures of protein−ligand complexes provide valuable insights into the interactions between proteins and small molecules. The statistical analysis of these structures has become an important tool in many different areas of research in the life sciences. Because of the large number of entries, the Protein Data Bank (PDB)[1] is the most important resource for experimentally determined structures of protein−ligand complexes. The structural data in the PDB is made available via different chemical file formats (PDB, mmCIF, PDBML/XML),[2] of which the PDB format[3] is the most common. PDB files contain element identities, three-dimensional coordinates, and connectivities for all atoms. However, unlike many other chemical file formats, this format does neither provide information about bond orders, formal charges, and aromaticity nor any kind of atom typing. Many cheminformatics methods and tools, however, depend on those and similar properties. Hence, when PDB files are supported as input, those properties have to be derived from the information provided by the file format. Although many current software packages include functionality to perceive molecular structures from three-dimensional coordinates, only a small number of these approaches has been published.[4−10]

The initial steps of all methods are similar to a certain extent. First, covalent bonds between atoms are identified by either using distance criteria or by simply relying on the connectivity data (CONECT entries) provided by the PDB format. In some approaches, this step is followed by a valence check during which spurious bonds arising from distorted geometries are removed. Subsequently, possible hybridizations for atoms are determined by analyzing bond lengths and bond angles. In the next step bond orders and atom types are assigned. Depending on the way these assignments are handled, the methods can be divided into two classes. Approaches from the first class determine bond orders independently of hybridization states, either by using the bond lengths directly or by matching of functional group patterns. This is often followed by an additional step during which inconsistencies in the assignments are handled. In methods from the second class, bond orders are derived directly from previously determined hybridization states using different bond localization routines.

We present a new method for the perception of molecular structures from three-dimensional atomic coordinates, which is based on the recently published NAOMI model.[11] Using its robust chemical description, the molecules are constructed in a hierarchical scoring approach. The first steps are based on the local geometry of each individual atom, whereas later steps include larger parts of the atom's environment to generate a correct chemical representation. This bottom-up approach has the advantage that it does not rely on definite assignments at early stages, for example, by assigning bond orders by torsion angles, or by matching of functional group patterns. In contrast to previously published methods, the final solution is selected from a list of potential candidate structures which are ranked using both confidence values for the atoms' geometry and

Received: July 30, 2012
Published: November 25, 2012

dx.doi.org/10.1021/ci300358c | *J. Chem. Inf. Model.* 2013, 53, 76−87

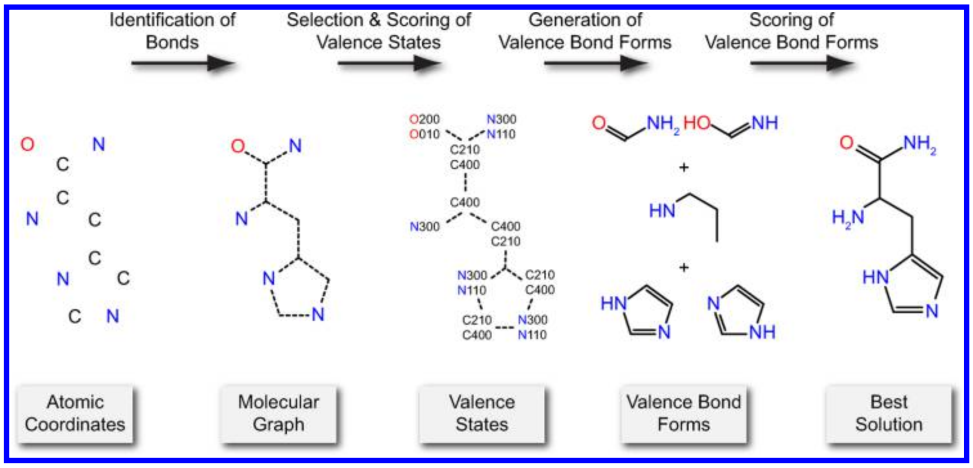


Figure 1. Schematic view of the workflow for the generation of molecules from three-dimensional coordinates.

chemical knowledge. This combination is the key to circumvent the shortcomings of other approaches, which either put too much focus on the provided coordinates or simply ignore them by using pattern-matching. The method's robustness and reliability are validated in different procedures by comparing reference molecules to the generated molecular structures. Furthermore, benchmark studies show its suitability for large scale applications.

■ METHODOLOGY

Overview. The aim of the presented method is the generation of both chemically valid and reasonable molecular structures from element identities and three-dimensional atomic coordinates. A molecule is considered chemically valid if a valence bond structure (Lewis structure) can be found, in which the valences of the atoms' elements are not violated. Not every possible valid valence bond form, however, provides a reasonable description of the molecule. On the one hand, geometric features, for example, interatomic distances and planar groups, must be reflected in the assigned bond orders. On the other hand, common standards for the representation of particular functional groups and resonance forms should be met. The last point is especially important since resonance forms and, depending on the quality of the provided coordinates, even tautomeric forms can not be deduced from geometry alone. For this purpose, we make use of the NAOMI model,[11] which has been successfully applied to the consistent conversion of chemical file formats. In this model, atoms are represented by three chemical descriptors, namely element, valence state, and atom type, which are assigned in three consecutive steps. Valence states represent valid bond order distributions for atoms in valence bond structures. They are defined by an element identity, the number of associated single, double, and triple bonds and a formal charge (e.g., N400+ for quaternary nitrogen atoms). As will be explained below, valence states can be used to generate valence bond forms if the atoms' connectivities are known. Atom types are derived from valence states and are thus independent of the input file format.

The perception of molecular structures from atomic coordinates is performed in four steps (see Figure 1). At first, covalent bonds are identified on the basis of interatomic distances. The second step comprises identification of possible valence states for each atom and scoring according to the atom's local environment. In the third step valence bond forms of the molecule are generated by enumerating valid combinations of valence states and their associated bond orders. These combinations are scored in the final step to determine the most appropriate valence bond representation of the molecule. The strategy adopted in our method is based on the opinion that the best possible compatibility between the perceived molecules and the provided coordinates should be sought. We believe, that the best way to do so is to build the molecular structure based on the atom's local geometries and use chemical knowledge only when either inconsistencies are encountered or ambiguities need to be resolved.

Identification of Bonds. To determine if a covalent bond exists between two atoms, the distance criterion originally proposed by Meng[4] is applied. A bond is created if

$$\delta_{\text{bond}} = r_{ij} - (R_i + R_j) < 0.4 \text{ Å} \tag{1}$$

where r_{ij} is the distance between the atoms i and j and R_i and R_j denote the covalent radii[12] of the atoms' corresponding elements. The high tolerance value of 0.4 Å in eq 1 ensures that no potential covalent bond is missed during the identification process. The softness of the criterion can, however, lead to an erroneous bond perception in case of distorted geometries. The resulting superfluous bonds give rise to two different types of chemical errors, which can readily occur at the same time. On the one hand, the atom's number of bonds may exceed the maximum valence of its associated element. On the other hand, distorted geometries may lead to the formation of incorrect cyclic structures (usually rings of size three or four). To deal with these errors, the bond perception is performed in several consecutive steps: (1) identification of bonds between non-hydrogen atoms, (2) valence check for all atoms and removal of superfluous bonds, (3) perception of the molecule's rings, (4) length check for all ring bonds and removal of superfluous bonds, and (5) identification of hydrogen bonds.

After the perception of all non-hydrogen bonds, each atom is checked for violations of its valence. This is done by comparing the number of identified bonds to the number of allowed bonds for its element. If a violation is encountered, long bonds (δ_{bond} > 0.1 Å) are removed in order of their lengths until either the valence is restored or all long bonds are eliminated. In case of short non-hydrogen bonds ($r_{ij} < 0.5 \cdot (R_i + R_j)$), the coordinates are considered incorrect and the molecule cannot be constructed. After the ring perception each ring is checked for long bonds (δ_{bond} > 0.1 Å). If such a ring is encountered, its longest bond is removed and the molecule's rings are

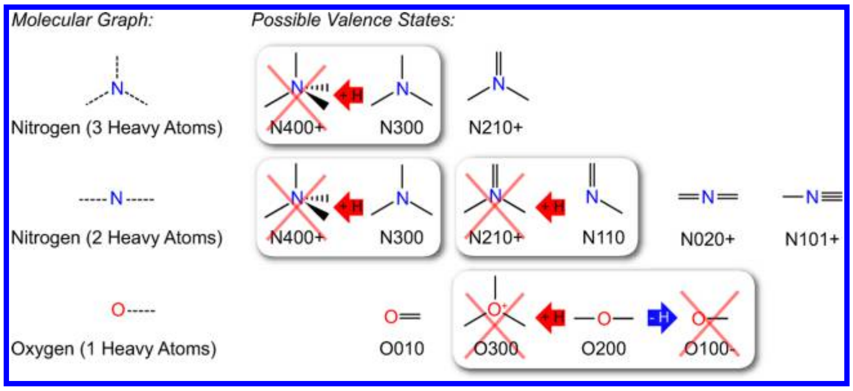


Figure 2. Examples for the selection of valence states. The crossed-out states are not selected since they can be deduced from the corresponding neutral states shown in the same box.

Table 1. Most Common Candidate Valence States for Typical Elements in Organic Molecules[a]

element	valence states						
hydrogen	H100						
carbon	C400	C210	C101	C020			
oxygen	O200	O010	O110+	O300+	O001+		
nitrogen	N300	N110	N210+	N400+	N020+	N101+	N001
phospohrous	P310	P300	P400+				
sulfur	S220	S210	S300+	S200	S110+	S010	S001+

[a] Valence states are represented as element symbol followed by the number of single, double, and triple bonds and the formal charge.

recalculated. This process is repeated until all long bonds in rings are eliminated. In contrast to non-hydrogen atoms, hydrogens are only allowed to have one bond and only the closest heavy atom needs to be identified. The hydrogen bond is created if the resulting bond is not short ($\Rightarrow r_{ij} \geq 0.5\cdot(R_i + R_j)$) and the heavy atom's valence is not violated. Otherwise the hydrogen atom is discarded.

Selection of Valence States. In the next step, suitable valence states are selected from a list of allowed states for the respective element for each atom. Since bond orders have not been assigned at this point and formal charges are usually not provided, the number of bonds from the previous step is the only criterion for this selection. Valence states are selected in two cases. First, if the valence state and the atom have an identical number of bonds. Second, if the atom's bond count is smaller, but the missing bonds can be saturated by hydrogens. Charged valence states are only considered if no corresponding neutral state exists or a formal charge has been specified for the atom. Examples for this identification procedure are shown in Figure 2.

In many cases, this results in an ambiguous assignment since multiple valence states may be compatible with a particular number of bonds. To deal with this ambiguity, all selected valence states are scored to determine the most appropriate choice as explained below. This score reflects the state's compatibility with the atom's local environment, which is characterized by the spatial distribution of the atom's neighbors and their respective element identities. The use of a predefined list of valid valence states is an important aspect of ensuring a molecule's chemical validity. Atoms with an invalid number of bonds can be easily identified by the fact that no candidate valence state has been found. This evidently applies to all cases, where the number of bonds exceeds the maximum allowed number for the respective element. In addition to that, it is also possible to identify atoms with unusual bond counts in case of higher row elements such as sulfur or phosphorus. A typical example would be a phosphate group that is missing two of its terminal oxygen atoms thus leaving the central phosphorus with only two covalent bonds. This constellation is rather unlikely in organic molecules and simply saturating the atom's valences by addition of hydrogens seems questionable in a chemical sense. If no candidate valence state for an atom can be found, the molecule is considered incorrect and cannot be constructed. The most common candidate valence states for typical elements in organic molecules are shown in Table 1.

Evaluation of Geometrical Parameters. The compatibility of valence states is mainly assessed on the basis of the atom's local geometry. For that purpose, several geometrical parameters g are evaluated and used to derive scores $G_p(g)$ for different chemical properties p, for example, bond orders. These scores are calculated according to the following scheme. For each property, a minimum and a maximum value are defined, which correspond to the scores of 0.0 and 1.0, respectively (see Table 2). Between the minimum and the maximum values a linear function is used.

The absolute value of the scalar **triple product** π of the normalized bond vectors connecting an atom and its neighbors

Table 2. Parameters for the Calculation of Scores $G_p(g)$ for Different Properties p[a]

property	parameter	minimum (0.0)	maximum (1.0)
$G_{planar}(\pi)$	π	≥0.6	≤0.15
$G_{linear}(\alpha)$	$\alpha[°]$	≤150	≥170
$G_{sp^2}(\alpha)$	$\alpha[°]$	≤114	≥118
$G_{single}(\delta)$	$\delta[Å]$	≤−0.1	≥−0.04
$G_{double}(\delta)$	$\delta[Å]$	≥−0.04	≤−0.1
$G_{triple}(\delta)$	$\delta[Å]$	≥−0.15	≤−0.25
$G_{planar}(\tau)$	$\tau[°]$	≥40	≤10

[a] Between the minimum and the maximum values a linear function is used.

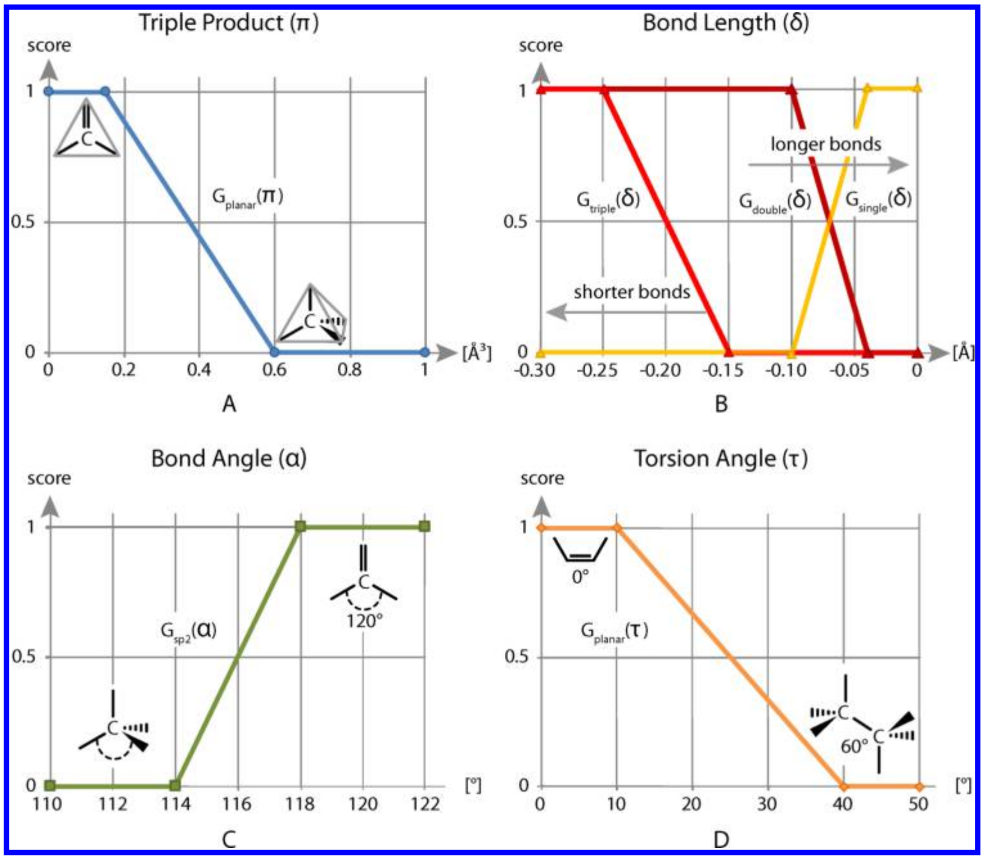


Figure 3. A: Score for the planarity of an atom using the triple product. B: Score for bond orders using the bond length. C: Score for an sp^2 hybridization on the basis of an atom's bond angle. D: Score for planarity using the largest torsion angle.

is a direct measure for its planarity ($G_{\text{planar}}(\pi)$) and can thus be used to distinguish sp^2 from sp^3 hybridizations. A triple product smaller than 0.15 indicates planarity, whereas a value larger than 0.6 (the triple product of an ideal tetrahedron is approximately 0.7) indicates the opposite. **Bond angles** α are used to determine the hybridization of an atom. They are especially important for the identification of linear geometries ($G_{\text{linear}}(\alpha)$), for example, in the presence of triple bonds. Because of the large difference to the bond angles of other hybridizations, sp hybridization can be easily distinguished. The smaller difference between the angles associated with sp^2 and sp^3 hybridizations makes the distinction between these cases rather difficult ($G_{\text{sp}^2}(\alpha)$). Scores for particular bond orders ($G_{\text{single}}(\delta)$, $G_{\text{double}}(\delta)$, $G_{\text{triple}}(\delta)$) are determined using the **bond length** δ which is calculated as described in eq 1. In the case of double bonds, the largest **torsion angle** τ at the respective bond is taken into consideration ($G_{\text{planar}}(\tau)$). Torsion angles can be used to check if the atoms surrounding the bond partners are coplanar, which is a precondition for double bonds. By taking torsion angles into account, invalid double bond assignments due to shortened interatomic distances can be avoided. Single bonds joining an aromatic ring with either an alkyl substituent or another aromatic ring are typical examples for this case. Although the bond length might be shortened, the torsion angle often clearly contradicts the double bond order. The torsion bond probability $G_{\text{double}}(\delta,\tau)$ is the product of $G_{\text{double}}(\delta)$ and $G_{\text{planar}}(\tau)$. For atoms in rings, **torsion angles** τ can be used to determine the planarity of the ring. In this case, only bonds in the same ring are included during the calculation of the largest torsion angle.

Probabilities of Hybridization States. The scores $G_{\text{p}}(g)$ are the basis for the calculation of probabilities for different hybridization states P_{hyb}. Since the number and kind of parameters used strongly depends on the atom's topology, each case is discussed separately.

For atoms with one bond, the bond length is the only available geometrical parameter.

$$P_{\text{sp}} = G_{\text{triple}}(\delta) \tag{2}$$

$$P_{\text{sp}^2} = G_{\text{double}}(\delta) - G_{\text{triple}}(\delta) \tag{3}$$

$$P_{\text{sp}^3} = G_{\text{single}}(\delta) \tag{4}$$

In this case the probabilities for the hybridization states correspond to the scores for the respective bond orders as described in eqs 2–4. Since sp hybridization is always associated with a linear geometry, the number of bonds at the atom's neighbor is checked. If the neighbor has more than two bonds this condition cannot be fulfilled and the value of P_{sp} is added to P_{sp^2} and then set to 0.0.

For atoms with two bonds, one bond angle and two bond lengths are available. The score for the presence of a double bond at the atom A_{double} is calculated as the sum of the torsion bond scores $G_{\text{double}}(\delta,\ \tau)$ of the atom's bonds, whereas its maximum value is limited to 1.0.

$$A_{\text{double}} = \min(1.0, \sum G_{\text{double}}(\delta,\ \tau)) \tag{5}$$

The sum in eq 5 is used to account for the limitations of valence bond structures. In delocalized systems, for example, aromatic rings, bonds can have lengths between the expected values of single and double bonds. In this case, the score for the presence of a double bond might be underestimated if only the larger of both values is considered. Because of the geometric restraints in small rings, we then distinguish two cases. For atoms in an acyclic environment or in large rings (at least eight atoms), the following probability scheme is used:

$$P_{\mathrm{sp}} = 2/3\cdot(G_{\mathrm{linear}}(\alpha) + 0.5\cdot A_{\mathrm{double}}) \tag{6}$$

$$P_{\mathrm{sp}^2} = \begin{cases} 1/2\cdot(1.0 - P_{\mathrm{sp}}) & \text{if} \quad P_{\mathrm{sp}} > 0.0 \\ 2/3\cdot(A_{\mathrm{double}} + 0.5\cdot G_{\mathrm{sp}^2}(\alpha)) & \text{else} \end{cases} \tag{7}$$

$$P_{\mathrm{sp}^3} = 1.0 - (P_{\mathrm{sp}^2} + P_{\mathrm{sp}}) \tag{8}$$

Since only the sp hybridization is compatible with a linear geometry, the bond angle has a higher weighting factor in the calculation of the associated probability in eq 6. For the probability of an sp^2 hybridization in eq 7 it is considered less reliable due to the small difference to the ideal value of the sp^3 hybridization. If the atom is part of a small ring (less than eight atoms), ring torsion angles can be used as an additional parameter to assess the planarity of the respective ring. Furthermore, a linear arrangement is extremely unlikely in these cases, so that only sp^2 and sp^3 hybridizations need to be considered. The probabilities and scores are adapted in the following way:

$$P_{\mathrm{sp}^2} = 2/5\cdot(A_{\mathrm{double}} + A_{\mathrm{planar}} + 0.5\cdot G_{\mathrm{sp}^2}(\alpha)) \tag{9}$$

$$P_{\mathrm{sp}^3} = 1.0 - P_{\mathrm{sp}^2} \tag{10}$$

Since bond angles in rings with a size smaller than six are strongly influenced by the strain of the cyclic arrangement, they are not a reliable measure for the atom's hybridization. In this case, the score is automatically set to 0.5 to indicate that no decision can be made. The planarity score A_{planar} in eq 9 for an atom is the minimum of the $G_{\mathrm{planar}}(\tau)$ (see Figure 3D) scores of each bond.

For atoms with three bonds, three bond angles, three bond lengths, and one triple product can be calculated. Since sp hybridization is not possible in this case, a decision between sp^2 and a sp^3 hybridization has to be made. For the calculation of the atom's angle score $A_{\mathrm{sp}^2}(\alpha)$ the mean bond angle $\overline{\alpha}$ is used.

$$P_{\mathrm{sp}^2} = 1/6\cdot(3\cdot G_{\mathrm{planar}}(\pi) + 2\cdot A_{\mathrm{double}} + A_{\mathrm{sp}^2}(\alpha)) \tag{11}$$

$$P_{\mathrm{sp}^3} = 1.0 - P_{\mathrm{sp}^2} \tag{12}$$

Again, the geometrical parameters are not considered equally reliable which is reflected in the different weighting factors in eq 11. The scoring of valence states for atoms with four or more bonds is solely based on scores for bond orders, and no probabilities for hybridizations need to be calculated for these cases.

Scoring of Valence States. The probabilities P_{hyb} from the previous step are used to calculate integer-based scores for all selected valence states of each atom. This score reflects the compatibility between the atom's local environment and the respective valence state and is used to identify the best suited state for an individual atom. Additionally, the absolute value of the score also provides a measure of confidence, which can be used to compare possible valence state assignments for different atoms. The scoring procedure makes use of the fact that valence states are not compatible with all hybridization states. In case of compatibility, the score S_{VS} is calculated using the probability P_{hyb} according to the following scheme:

$$S_{\mathrm{VS}} = \begin{cases} 1 & \text{if} \quad P_{\mathrm{hyb}} < 0.6 \\ \lfloor P_{\mathrm{hyb}}\cdot c + 0.5 \rfloor & \text{else} \end{cases} \tag{13}$$

The confidence factor c in eq 13 determines the maximum value of the score and depends on the topology of the respective atom (see Table 3). The values are based on the number of geometrical parameters available for the calculation of the probabilities P_{hyb}. A single bond length, for example, is not well suited to reliably distinguish between hybridizations, since even small geometrical distortions may cause the bond order perception to fail. This lack of reliability is reflected in a small confidence factor of 2.0 for atoms with one bond. The integer-based scheme ensures that only those valence states which are clearly favored by the atom's local geometry receive scores larger than one. This prevents the elimination of valence states based on small geometrical differences.

Table 3. Confidence Values for Different Topologies

topology	confidence c
1 bond	2.0
2 bonds(acyclic)	3.0
2 bonds(cyclic)	4.0
≥3 bonds	5.0

If the compatibility between the selected valence states and their associated hybridization states is mutually exclusive, the scoring procedure is straightforward. Because of the limitation of valence bond structures, this is, however, not always the case (see Figure 4 for examples). On the one hand, there are atoms which are represented by the same valence state but have different hybridizations, such as nitrogens in amines and amides. These cases are handled by assigning the largest score obtained for all compatible hybridizations to the respective valence state. On the other hand, some atoms are not sufficiently represented by a single valence state such as oxygens in a carboxylate group. In this case both compatible valence states receive identical scores. Examples for the scoring procedure are shown in Figure 5.

The calculation of scores for atoms with four or more bonds can in most cases be avoided due to the fact that there is only one suitable valence state. If this is not the case, the multiple bond score A_{double} introduced in eq 5 is used in place of P_{hyb} to calculate the score for all selected valence states. This is always sufficient to distinguish between the alternatives.

In some cases, it is beneficial to remove valence states from the list of candidates if their associated hybridization is not compatible with the atom's local geometry ($P_{\mathrm{hyb}} = 0.0$). These valence states will not be considered during the generation of valence state forms, which in turn reduces the complexity of the next steps. Since distorted geometries could easily lead to the premature exclusion of relevant valence states, this is only done in two rather unambiguous cases. First, if the corresponding valence state is only compatible with an sp hybridization and second if the atom has three bonds.

Distorted geometries can also result in incorrect scores which will eventually lead to undesired valence bond structures. This is especially true if atoms with only one bond are involved since the resulting assignment cannot be corrected by the valence states of the surrounding atoms. To avoid these errors, valence

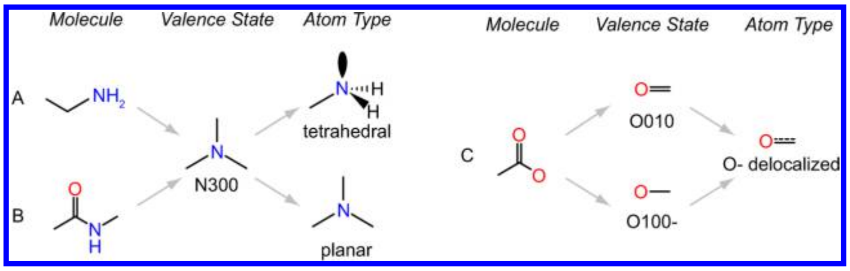


Figure 4. Limitations of valence bond structures: (a) Nitrogens in amides and amines have the same valence states but different geometries. (b) Oxygens in carboxylates have different valence states but have the same bond length.

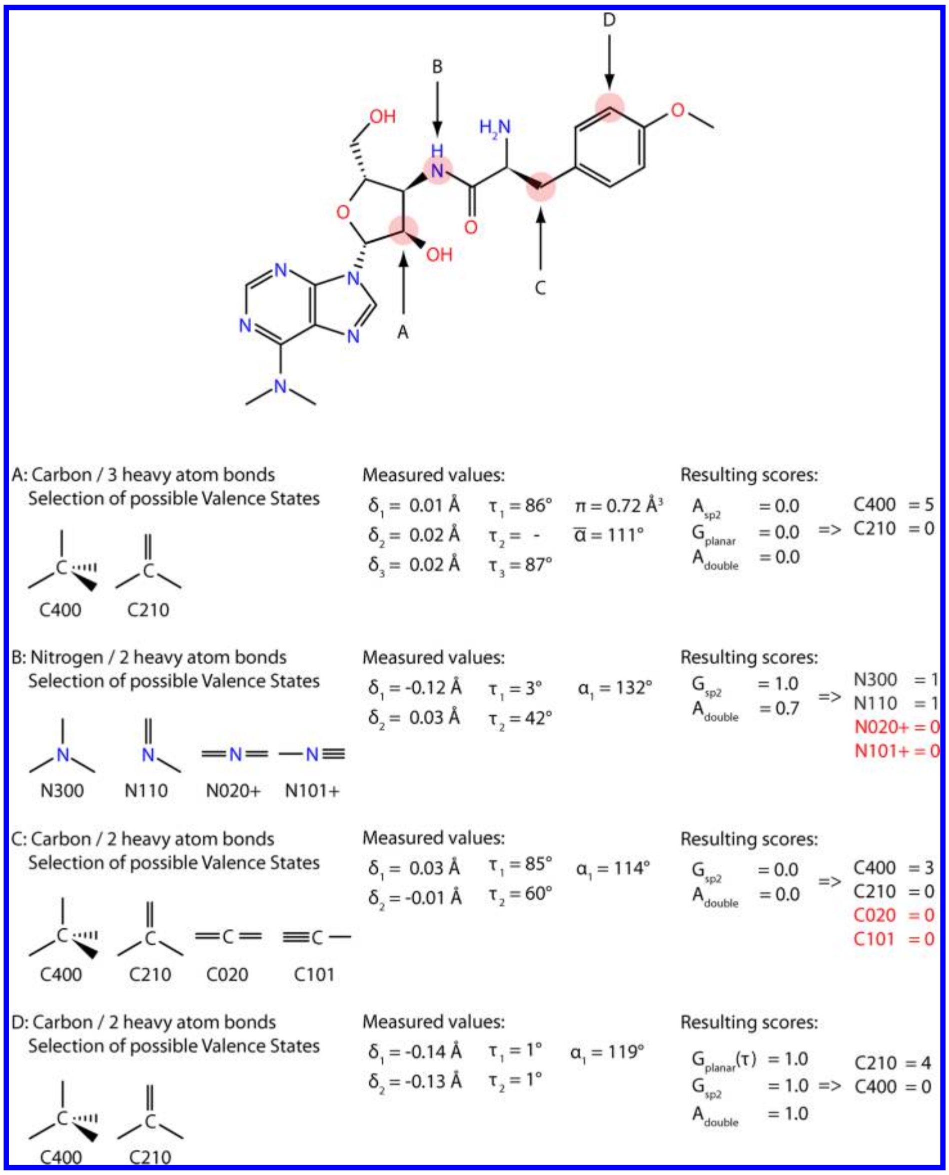


Figure 5. Examples for the valence state scoring procedure. Relevant geometrical parameters are triple products, bond lengths, bond angles and torsion angles. If bond angles and bond lengths do not indicate a linear geometry, valence states associated with a linear geometry are excluded (marked in red). For the atoms A, C, and D the geometrical parameters clearly support one of the valence states. In the case of atom B, there is no clear preference and two valence states are equally probable.

state scores of atoms that are part of specific substructures are increased by +2 (see Figure 6). These resulting scores are, however, not high enough to change the assignment in case of a perfect geometric compatibility.

The purpose of the described procedure is to provide reliable scores which can be used to identify the best valence state representation of the molecule. Due to the differing quality of available input data, this must also apply if the provided coordinates are of poor quality. As mentioned above, the scores are not only used to find the best choice for an individual atom but also to compare assignments between atoms. This means that valence states with higher scores have a stronger influence

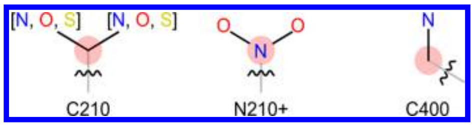


Figure 6. Additional scores of +2 are assigned to valence states for atoms (marked with red spheres) in specific substructures. The number of bonds at the atoms corresponds to the number of bonds identified during the bond perception.

on the resulting valence state form. The score does not only depend on the number of available parameters at the respective atom, but also on their consistency. This is assessed by the individual evaluation of the geometrical parameters during the calculation of the probabilities P_{hyb}. A value of 1.0 is only possible if all geometrical parameters are consistent, which in turn results in low scores for atoms with inconsistent local geometries.

Generation of Valence Bond Forms. In the next step, chemically valid valence bond representations of the molecule are generated by assigning valence states to all atoms and bond orders to all bonds. A combination of valence states is valid if a bond order distribution can be generated which is in accordance with the valence states of the atoms. The score of such a combination is calculated as the sum of the scores of the valence states from the previous step. The best combination can be identified by enumerating all valid combinations with a maximum score. For the enumeration a branch and bound algorithm with a depth-first search strategy is used. Prior to the enumeration, the list of valence states for each atom is checked for cases where only one valence state is remaining. This state is assigned directly and the orders of the adjacent bonds are adapted accordingly. Afterward, the molecule is partitioned into zones containing atoms connected by bonds with unassigned bond orders. The individual processing of each zone further decreases the number of possible combinations. If a single best scored combination for a zone exists, it is selected. Otherwise, combinations with equal scores are ranked using additional geometrical and chemical criteria as described below.

Scoring of Valence Bond Forms. Each combination of valence states generated in the last step is a valid valence bond form (in the sense that no valences are violated) and is also compatible with the local geometry of the atoms. This does, however, not necessarily imply that each form provides a reasonable description of the molecule. On the one hand, discrepancies between the assigned bond orders and the actual bond lengths might exist, which could not be resolved during the atom based valence state scoring procedure. On the other hand, the combination might contain unusual representations of functional groups or conjugated systems, which could not be excluded using geometrical parameters alone. Hence, an additional scoring scheme, which makes explicit use of the assigned bond orders, is applied to distinguish reasonable from undesired valence bond forms. In contrast to the previous steps, where geometrical parameters had a high priority, this step focuses mainly on chemical aspects.

Prior to the scoring procedure, valence states and bond orders are assigned if they are identical in all generated valence bond forms. Afterward, the molecule is again partitioned into zones containing atoms connected by bonds with unassigned bond orders. Then, substructures (see Figure 7) including at least one of the unassigned atoms are identified in each of the remaining valence bond forms. These substructures correspond to preferred representations of functional groups and for each match a score of +1 is assigned to the respective form. If an unassigned atom is part of a ring with a size smaller than eight, Hueckel's rule is applied to assess its aromaticity. Valence bond forms receive a score of +1 for each ring, where the rule is fulfilled. It must be stressed that our approach does not favor particular functional groups or aromatic rings in general but only if the geometrical parameters were not sufficient to resolve the structure.

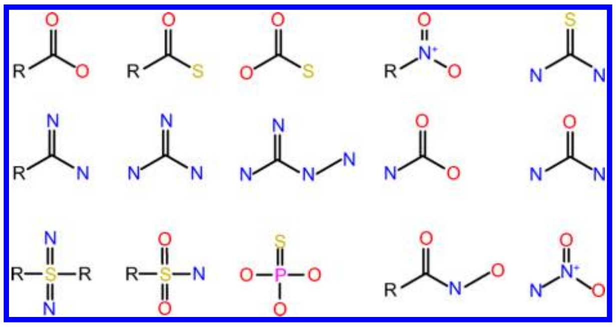


Figure 7. Substructures representing favored representations of particular functional groups in valence bond structures. The R represents both carbon and hydrogen.

If a bond with an unassigned bond order is part of a substructure or ring which has been scored in the previous step, no further scoring is performed. Otherwise, $G_{\text{order}}(\delta)$ from Table 2 is used to determine if the current bond order is compatible with the calculated bond length. In the case of double bonds, $G_{\text{double}}(\delta,\tau)$ is used. If the respective value exceeds a threshold of 0.7 a score of +1 is assigned to the valence bond form. Hence, solutions in which bond lengths do not correspond to the assigned bond orders receive lower scores. If the bond is also part of a ring with less than eight atoms, $G_{\text{planar}}(\tau)$ is used as an additional parameter to assess the bond's planarity. A score of +1 is assigned if either $G_{\text{planar}}(\tau)$ is smaller than 0.3 (planar geometry) for a double bond or $G_{\text{planar}}(\tau)$ is larger than 0.7 (ring is not planar) for a single bond.

Again, only the solutions with the largest scores are kept. If there are still multiple solutions left, they are considered equivalent and a canonization scheme is used to choose a unique form for each zone. Since a detailed explanation of the canonization algorithm extends the scope of this publication, only a brief description of the general idea will be given. The atoms of the respective zone are ordered in a procedure similar to the CANON algorithm[13] used for the generation of USMILES. The zone is then processed atom by atom according to this newly generated order. At each step the respective solutions are sorted by the valence states (using ids as sorting criterion) of the particular atom and all solutions with lower ranks are eliminated. This process is repeated until only one solution remains. Obviously, it is also possible to omit the canonization and use the solutions for each zone to enumerate all equivalent valence bond forms of the molecule.

■ RESULTS AND DISCUSSION

Validation with Curated Structures. In a first validation procedure we tested if our method was able to generate the expected valence bond structures for small molecules from different PDB entries. The success was verified by comparison of the resulting molecules to manually curated reference structures provided as USMILES.[13] Small molecules were extracted from PDB entries used in the studies of Hendlich[6] and Labute.[7] Because of its importance in the field of cheminformatics, we also included the ligands from the PDB entries of the Astex Diverse Set.[14] The complete validation set consists of 563 molecules from 363 PDB entries. Both PDB entries and SMILES files for the respective compounds are provided as Supporting Information. Table 4 lists the PDB entries and the component names of the ligands for which our method failed to generate the expected structure. Five of these examples are shown together with the reference structures taken from the respective publications in Figure 8.

Table 4. PDB IDs and Component Names of All Molecules for Which Our Method Did Not Generate the Expected Structure

Labute[7]	Hendlich[6]	Astex[14]
2R04 (W71)	1MIO (CFM)	1G9V (HEM)
3FX2 (FMN)	1PMP (OLA)	1Q4G (HEM)
5TLN (BAN)	6RSA (UVC)	
8XIA (XLS)		

The dihydro-oxazol ring of ligand W71 from 2R04 (see Figure 8A) is perceived as oxazol. Because of a short bond of C4A to the nitrogen atom and the planarity of the five-membered ring, the valence state C210 (which is compatible with an sp^2 hybridization) receives a higher score. This eventually leads to a structure including an aromatic ring. One of the hydroxy groups of the flavin mononucleotide ligand FMN from 3FX2 (see Figure 8B) is interpreted as a carbonyl group. In this case the valence state C210 is favored due to the trigonal planar geometry of C2′. The same also applies to the α carbon CA2 in BAN from 5TLN (see Figure 8E). The carbonyl group of the molecule XLS from 8XIA (see Figure 8C) is interpreted as a hydroxy group because of the tetrahedral geometry at C2. The double bonds of the olefinic moiety of OLA (1PMP) (see Figure 8D) and of one of the vinylic groups in HEM (1G9V, 1Q4G) are perceived as single bonds due to the bond lengths and associated bond angles.

Our method was able to generate the correct structure in 98% of the cases. All observed differences were caused by strong deviations from the expected molecular geometries. The valence bond forms generated by our method are, however, equally reasonable in a chemical sense and also in agreement with the supplied atomic coordinates. Only in the case of BAN the generated structure does not correspond to the tautomeric form which would be expected for the isolated compound with respect to the hydroxamic acid group. The molecular geometry may, however, be influenced by the interactions with a metal atom in the protein−ligand complex. The PDB entry CFM contains an Fe−Mo−S cluster, for which our method does not produce a valence bond form but isolated atoms. Since valence bond forms are not well suited to describe metal clusters, we do not consider this a perception error, but think it should be mentioned at this point. The same is true for the vanadate in 6RSA in which no bonds between the oxygens and the vanadium atom are formed. The uridine molecule, however, is perceived correctly.

Comparison with Other Methods. To compare our results with those of other existing methods, we used the tools I-interpret,[15] fconv[16] and MOE[17] to generate molecules for the above-mentioned 363 PDB entries. This was done by first converting the entries from PDB to SDF (since fconv does not support sdf as output format, mol2 was chosen in this case) and then using the converted file as input for the comparison to the reference structures. The results are summarized in Table 5. Since our method will be part of the NAOMI converter, it is referred to as NAOMI in the table. The comparison to the reference structures was done using the NAOMI framework. Since all files (PDB input, SDF/MOL2 files from different tools, SMILES for comparison) are supplied as Supporting

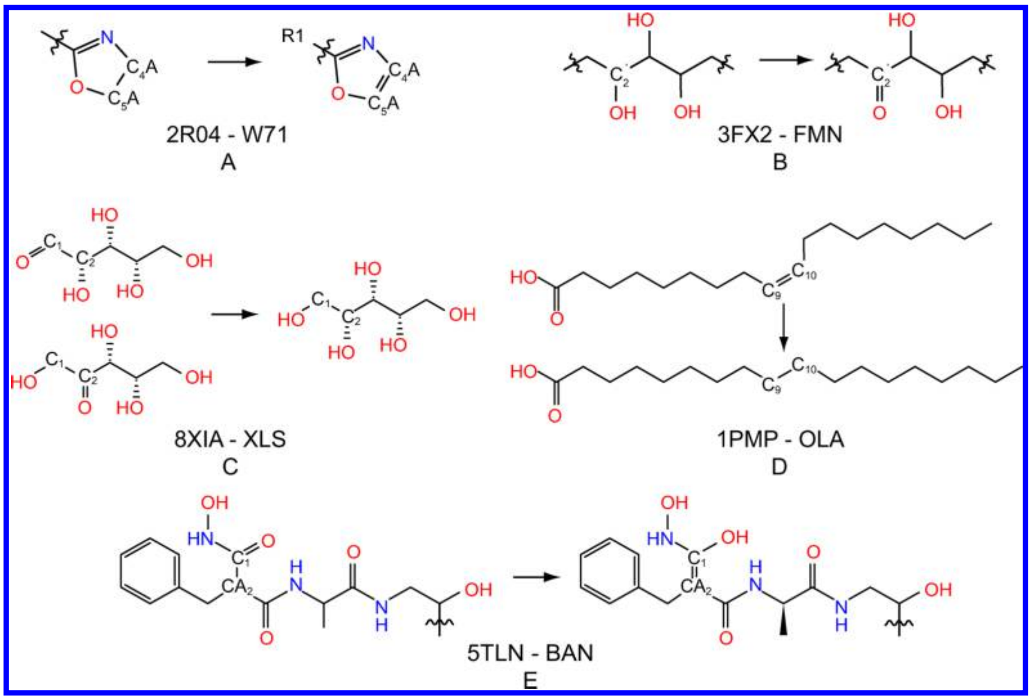
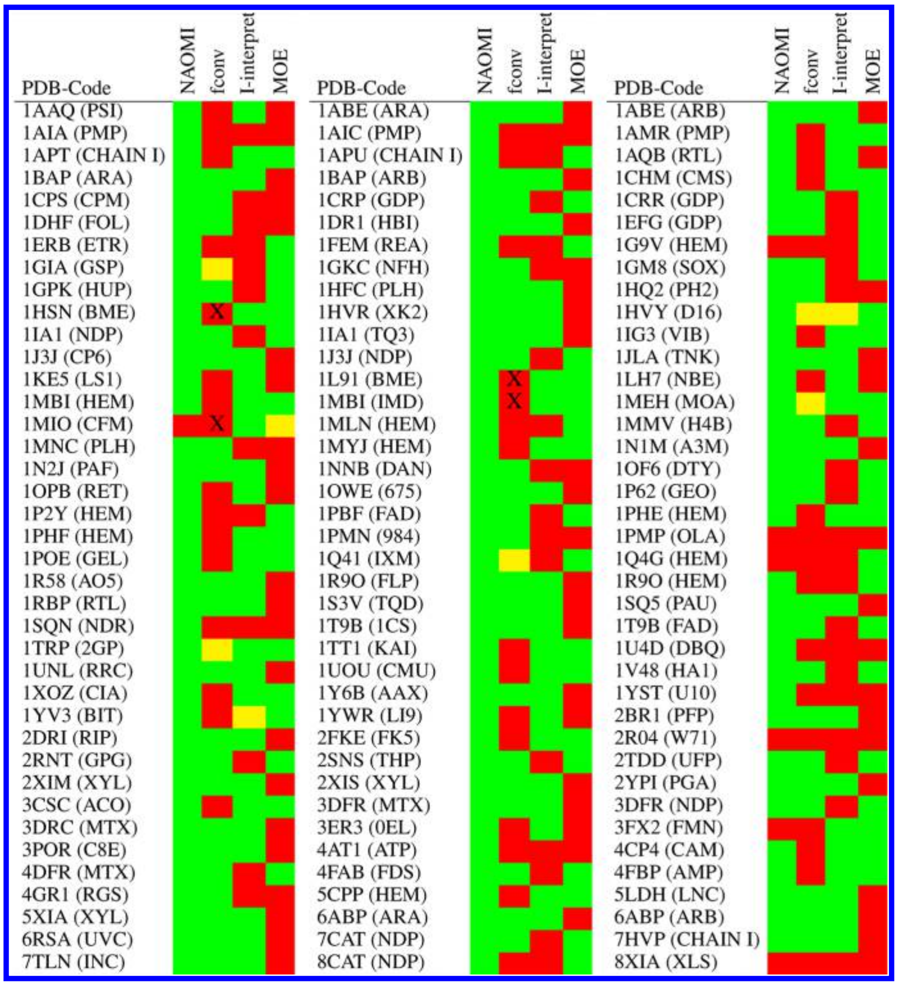


Figure 8. Five of the nine molecules for which our method did not generate the expected structure. The expected results are shown on the left side of the arrow, the results of our method on the right. The names from the PDB files are listed for all atoms for which incorrect valence states were identified.

Table 5. Results of the Generation of Molecules from the 363 PDB Entries Using Different Tools[a]



PDB-Code	NAOMI	fconv	I-interpret	MOE	PDB-Code	NAOMI	fconv	I-interpret	MOE	PDB-Code	NAOMI	fconv	I-interpret	MOE
1AAQ (PSI)	🟩	🟥	🟩	🟥	1ABE (ARA)	🟩	🟩	🟩	🟥	1ABE (ARB)	🟩	🟩	🟩	🟥
1AIA (PMP)	🟩	🟥	🟥	🟥	1AIC (PMP)	🟩	🟥	🟥	🟥	1AMR (PMP)	🟩	🟥	🟩	🟩
1APT (CHAIN I)	🟩	🟥	🟩	🟩	1APU (CHAIN I)	🟩	🟥	🟥	🟩	1AQB (RTL)	🟩	🟥	🟩	🟥
1BAP (ARA)	🟩	🟩	🟩	🟥	1BAP (ARB)	🟩	🟩	🟩	🟥	1CHM (CMS)	🟩	🟥	🟩	🟩
1CPS (CPM)	🟩	🟩	🟥	🟥	1CRP (GDP)	🟩	🟩	🟥	🟩	1CRR (GDP)	🟩	🟩	🟥	🟩
1DHF (FOL)	🟩	🟩	🟥	🟥	1DR1 (HBI)	🟩	🟩	🟩	🟥	1EFG (GDP)	🟩	🟩	🟥	🟩
1ERB (ETR)	🟩	🟥	🟥	🟩	1FEM (REA)	🟩	🟥	🟥	🟩	1G9V (HEM)	🟥	🟥	🟥	🟩
1GIA (GSP)	🟩	🟨	🟥	🟩	1GKC (NFH)	🟩	🟩	🟥	🟥	1GM8 (SOX)	🟩	🟩	🟥	🟩
1GPK (HUP)	🟩	🟩	🟥	🟩	1HFC (PLH)	🟩	🟩	🟩	🟥	1HQ2 (PH2)	🟩	🟩	🟥	🟥
1HSN (BME)	🟩	🟥 X	🟩	🟩	1HVR (XK2)	🟩	🟩	🟩	🟥	1HVY (D16)	🟩	🟨	🟨	🟩
1IA1 (NDP)	🟩	🟩	🟥	🟩	1IA1 (TQ3)	🟩	🟩	🟩	🟥	1IG3 (VIB)	🟩	🟥	🟩	🟩
1J3J (CP6)	🟩	🟩	🟩	🟥	1J3J (NDP)	🟩	🟩	🟥	🟩	1JLA (TNK)	🟩	🟩	🟩	🟥
1KE5 (LS1)	🟩	🟥	🟩	🟥	1L91 (BME)	🟩	🟥 X	🟩	🟩	1LH7 (NBE)	🟩	🟥	🟩	🟥
1MBI (HEM)	🟩	🟥	🟩	🟩	1MBI (IMD)	🟩	🟥 X	🟩	🟩	1MEH (MOA)	🟩	🟨	🟩	🟩
1MIO (CFM)	🟥	🟥 X	🟩	🟨	1MLN (HEM)	🟩	🟥	🟥	🟩	1MMV (H4B)	🟩	🟩	🟥	🟩
1MNC (PLH)	🟩	🟩	🟥	🟥	1MYJ (HEM)	🟩	🟥	🟩	🟩	1N1M (A3M)	🟩	🟩	🟩	🟥
1N2J (PAF)	🟩	🟩	🟩	🟥	1NNB (DAN)	🟩	🟩	🟥	🟥	1OF6 (DTY)	🟩	🟩	🟥	🟩
1OPB (RET)	🟩	🟥	🟩	🟥	1OWE (675)	🟩	🟩	🟩	🟥	1P62 (GEO)	🟩	🟩	🟥	🟩
1P2Y (HEM)	🟩	🟥	🟥	🟩	1PBF (FAD)	🟩	🟩	🟥	🟩	1PHE (HEM)	🟩	🟥	🟩	🟩
1PHF (HEM)	🟩	🟥	🟩	🟩	1PMN (984)	🟩	🟩	🟥	🟥	1PMP (OLA)	🟥	🟥	🟥	🟥
1POE (GEL)	🟩	🟥	🟩	🟩	1Q41 (IXM)	🟩	🟨	🟥	🟩	1Q4G (HEM)	🟥	🟥	🟥	🟩
1R58 (AO5)	🟩	🟩	🟩	🟥	1R9O (FLP)	🟩	🟩	🟩	🟥	1R9O (HEM)	🟩	🟥	🟥	🟩
1RBP (RTL)	🟩	🟩	🟩	🟥	1S3V (TQD)	🟩	🟩	🟩	🟥	1SQ5 (PAU)	🟩	🟩	🟩	🟥
1SQN (NDR)	🟩	🟥	🟥	🟥	1T9B (1CS)	🟩	🟩	🟩	🟥	1T9B (FAD)	🟩	🟩	🟥	🟩
1TRP (2GP)	🟩	🟨	🟩	🟩	1TT1 (KAI)	🟩	🟥	🟩	🟩	1U4D (DBQ)	🟩	🟥	🟥	🟥
1UNL (RRC)	🟩	🟩	🟩	🟥	1UOU (CMU)	🟩	🟥	🟩	🟩	1V48 (HA1)	🟩	🟩	🟥	🟩
1XOZ (CIA)	🟩	🟥	🟩	🟩	1Y6B (AAX)	🟩	🟩	🟩	🟥	1YST (U10)	🟩	🟥	🟥	🟥
1YV3 (BIT)	🟩	🟥	🟨	🟩	1YWR (LI9)	🟩	🟥	🟩	🟥	2BR1 (PFP)	🟩	🟩	🟩	🟥
2DRI (RIP)	🟩	🟩	🟩	🟥	2FKE (FK5)	🟩	🟥	🟩	🟩	2R04 (W71)	🟥	🟥	🟥	🟥
2RNT (GPG)	🟩	🟩	🟥	🟩	2SNS (THP)	🟩	🟩	🟥	🟩	2TDD (UFP)	🟩	🟩	🟥	🟩
2XIM (XYL)	🟩	🟩	🟩	🟥	2XIS (XYL)	🟩	🟩	🟩	🟥	2YPI (PGA)	🟩	🟩	🟩	🟥
3CSC (ACO)	🟩	🟥	🟩	🟩	3DFR (MTX)	🟩	🟩	🟩	🟥	3DFR (NDP)	🟩	🟩	🟥	🟩
3DRC (MTX)	🟩	🟩	🟩	🟥	3ER3 (0EL)	🟩	🟥	🟩	🟥	3FX2 (FMN)	🟥	🟥	🟩	🟩
3POR (C8E)	🟩	🟩	🟩	🟥	4AT1 (ATP)	🟩	🟥	🟥	🟥	4CP4 (CAM)	🟩	🟥	🟩	🟩
4DFR (MTX)	🟩	🟩	🟥	🟩	4FAB (FDS)	🟩	🟩	🟥	🟩	4FBP (AMP)	🟩	🟥	🟩	🟩
4GR1 (RGS)	🟩	🟩	🟥	🟥	5CPP (HEM)	🟩	🟥	🟩	🟩	5LDH (LNC)	🟩	🟩	🟩	🟥
5XIA (XYL)	🟩	🟩	🟩	🟥	6ABP (ARA)	🟩	🟩	🟩	🟥	6ABP (ARB)	🟩	🟩	🟩	🟥
6RSA (UVC)	🟩	🟩	🟩	🟥	7CAT (NDP)	🟩	🟩	🟥	🟩	7HVP (CHAIN I)	🟩	🟩	🟩	🟥
7TLN (INC)	🟩	🟩	🟩	🟥	8CAT (NDP)	🟩	🟥	🟥	🟩	8XIA (XLS)	🟥	🟥	🟥	🟥

[a] The colors represent the quality of the resulting structures. Green cells: Correct structure. Yellow cells: Suboptimal structure. Red cells: Structure substantially differing from reference. X: No structure generated.

Information, the comparison can be carried out using other tools with the same functionality. The differences between the generated molecules and the references can be divided into two categories. First, there are molecules for which hybridization states or bond orders have been differently assigned. All of these differences are caused by deviations from the expected geometries and are thus directly linked to the quality of the respective coordinates. Second, there are molecules with unusual or even chemically unreasonable resonance or tautomeric forms. Although these differences are not wrong considering the molecule's geometry, they deviate from conventions concerning the representation of particular substructures. Depending on the gravity of these deviations, the solutions are either considered invalid or simply not optimal. Examples for both cases are shown in Figure 9.

Table 5 shows that many differences appearing with other tools are avoided by our method. Incorrect perceptions because of geometrical distortions are often prevented by considering all aspects of an atom's environment. The confidence values for valence states are derived from multiple geometrical parameters so that the assignment has a certain stability against small geometrical distortions. This is a considerable advantage over methods which rely on definite assignments based on particular geometrical parameters. By considering the confidence values of surrounding atoms during the generation of valence bond structures even strong distortions can be compensated in some cases. The explicit inclusion of chemical knowledge in the last step of the workflow helps to reliably resolve the remaining ambiguities. Errors concerning the representation of molecules typically occur with methods that put too much emphasis on the evaluation of the geometrical parameters during the generation of valence bond forms. One has to keep in mind that localized bond orders are only an approximation and do

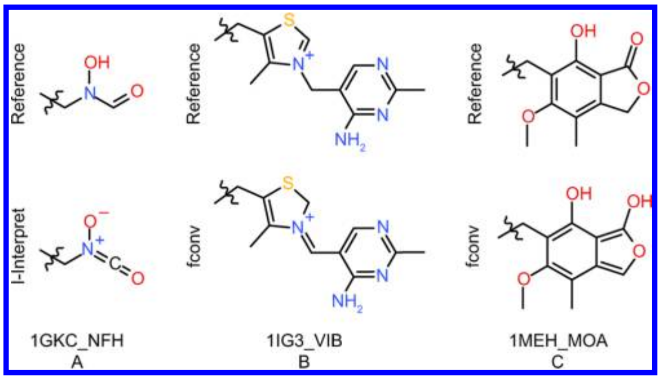


Figure 9. Comparison of reference structures and perceived structures generated by other tools. The structures A and B are classified as errors, whereas structure C is classified as not optimal.

not have to strictly adhere to molecular geometries. By scoring multiple alternative structures using a combination of chemical and geometrical criteria, our method is able to generate molecules that are both in agreement with the atomic coordinates and chemically reasonable.

Validation with Complete Ligand Expo Data Set. The main purpose of our method is the automatic generation of reasonable molecular representations for large data sets. To show that our method is both, efficient and robust, we applied it to all entries of the Ligand Expo data set[18] in PDB format and analyzed the results in terms of runtime and quality. The generated structures were compared to the respective molecules in the SDF format, which are also provided on the Ligand Expo Web site.[19] Again, USMILES served as a basis for the comparison. Since the NAOMI model does not support covalently bound metal atoms, all metal bonds were ignored and only the largest resulting component was used. Additionally, monatomic entries were skipped, since the ionization state of single atoms can not be deduced without knowledge of the environment. Empty entries and entries with multiple disconnected components were also ignored, since this usually indicates missing atoms. Some entries were rejected due to unusually small distances between atoms (coordinate errors). The results of this procedure are summarized in Table 6.

Table 6. Results of the Analysis of the 602704 Entries in the Ligand Expo Data Set for Both SDF and PDB

	SDF	PDB
no. total	602704	602704
mo. format errors	0	3015
no. empty entries	7688	7678
no. monatomic entries	241002	239452
no. disconnected entries	10254	10193
no. coordinate errors	499	939
no. converted entries	343261	341427
no. compared entries	334121	

Both data sets initially contained 602704 entries, of which 334121 (55.4%) were eventually used for comparison. To avoid inconsistencies concerning ionization states, all molecules were neutralized in advance (see Figure 10). In 91.7% (306341) of the cases identical valence bond structures were found. The reasons for the observed 27780 differences are quite diverse, as shown shown in Table 7.

Table 7. Analysis of the Reasons for Different Valence Bond Structures for the 334121 Compared Entries of the Ligand Expo Data Set[a]

	entries	% of data set	% of differences
no. different valence bond form	27780	8.3	100
no. different tautomeric form	10012	3.0	36.0
no. different oxidation state	810	0.2	2.9
no. different bond order	10349	3.1	37.3
no. different terminal bond order	6063	1.8	21.8
no. small molecule	3523	1.1	12.7

[a] Molecules are considered small if they have less than 8 heavy atoms.

In 10012 (36.0%) of the cases, a different tautomeric form of the molecule was generated. Tautomeric forms can often not be distinguished on the basis of the provided coordinates and multiple solutions are equally acceptable. As described above these cases are handled by a canonization procedure, so that different tautomeric forms do not indicate perception errors but rather different default representations. Typical examples for substructures with equivalent tautomeric states are substituted imidazoles, pyrimidones, and guanidinium groups. 810 (2.9%) of the differences were due to different oxidation states of particular heterocyclic compounds such as NAD/NADP. As with tautomers, these states can not be reliably distinguished on the basis of atomic coordinates, especially in entries with low resolution. Therefore, these cases are also not considered perception errors meaning that 94.9% of the results are essentially identical.

The remaining 16958 entries were further investigated in order to determine the reason for the incorrect perception. These entries correspond to 2341 different components, of which the 20 with the highest counts are shown in Table 8. Evidently, 22.8% (3864) of the differences are caused by only 1% of the components. These entries will be used for the discussion of specific problems encountered with the LigandExpo data set.

The errors associated with HEM are almost exclusively caused by the vinylic double bonds. As discussed above, the number of available geometrical parameters for the determination of bond orders for terminal bonds is small and makes the perception less stable with respect to deviations from ideal geometries. PGV, BCR, PEK, PEV, and OLC are molecules with long aliphatic chains and a specific number of double bonds. In many entries there is a considerable disagreement between our method and the LigandExpo references concerning both the presence and position of these double bonds. We have encountered numerous examples where we did not even find a single shortened bond length in the molecule although a double bond was present in the LigandExpo structure. Many of the incorrect perceptions concerning FAD, NAD, and UMP are caused by strong geometrical distortions of the respective aromatic rings. In some cases torsion angles that reach up to 40° are encountered in these usually completely planar structures. In case of CYC, BLA, and MDO exocyclic carbon−carbon double bonds at five-membered aromatic heterocyclic are interpreted as single bonds. These assignments were in all cases a result of an unambiguous single bond length at the respective bond. The difference from the entries ACB, MLE, and MYR are caused by the specific way covalently bound compounds are handled in the PDB format. If a molecule is bound to a residue of a protein or nucleic acid, the atom involved in this bond is usually assigned to the residue.

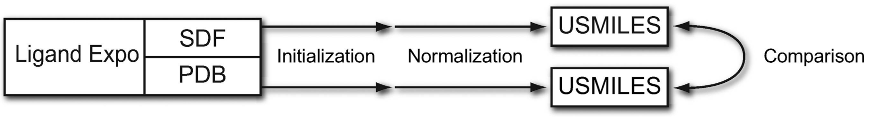


Figure 10. Scheme for the comparison of molecules from the Ligand Expo data set. Generated molecules from the PDB format are compared to the respective structures from the SDF format.

Table 8. PDB Component Names and Numbers of Errors for Those Molecules for Which the Most Errors Occured

name	no. errors	name	no. errors	name	no. errors	name	no. errors	name	no. errors
HEM	1794	PGV	194	CYC	187	BCR	182	LLP	164
ACB	124	FAD	124	PEK	120	PEV	102	MLE	84
PSO	124	BLA	83	1MA	81	MYR	80	7MG	80
OLC	79	MDO	77	NAD	77	PDU	76	UMP	73

This means that the compound in the entry does not represent an isolated molecule and that necessary information is missing. These errors can often be avoided when the complete PDB entry including the protein environment is used. The reasons for the differences encountered for PSO are quite similar. The psoralen is also covalently bound to a nucleotide but in this case no atoms from the initial component are missing. This connection is, nevertheless, reflected in the coordinates by a change of hybridization geometries for the carbon atoms in the five-membered ring. Since the molecule contained in the LigandExpo data set is an isolated psoralen, the different perception is not surprising. In case of LLP, PDU, and 7MG the structures provided by the LigandExpo data set seem to be wrong (see Figure 11).

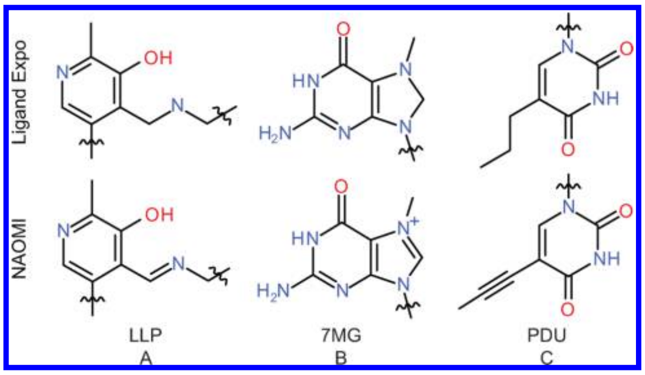


Figure 11. Comparison of inconsistent structures from the LigandExpo data set to those generated by NAOMI.

The compound LLP represents a lysine residue covalently linked to a pyridoxal phosphate via an imine group. This double bond is not present in any of the structures from LigandExpo although it is reflected by a short bond length in the coordinates. The name for the compound PDU on the LigandExpo Web site is 5(1-propynyl)-2′-deoxyuridine-5-monophosphate which indicates the presence of a triple bond. This is also confirmed by an analysis of the molecule's geometry. This triple bond is, however, not present in the reference structure. 7MG is supposed to be 7*N*-methyl-guanosine-5′-monophosphate, a molecule with a charged five-membered heterocycle which is generated by our method. The structure found in the LigandExpo data set, however, has a carbon atom with an sp^3 hybridization in the five-membered ring.

We think that these examples are sufficient to provide a general overview of the reasons for the observed differences. A special case worth mentioning are molecules with fewer than eight heavy atoms, such as solvents and auxiliary agents. Because of the extreme deviations from ideal geometries, these entries can often not be handled on the basis of atomic coordinates alone. We believe that in some cases these molecules were of minor interest to the researchers and less care was taken during the structure determination process. When interpreting the results of the comparison one has to keep in mind that our method solely relies on the atomic coordinates provided by the file format. The reference molecules in the Ligand Expo data set are, however, derived from various inputs. In particular, this includes information about the components provided by the crystallographers. This means that the provided coordinates are not necessarily in perfect agreement with the structures present in the data set. In the end 10349 (61.0%) of the 16958 remaining entries differ by only one bond order and the respective bond is terminal in 6063 (35.8%) of these cases. This shows that the generated structures, even if they are not identical, are generally in good agreement for the larger part of the molecules.

Runtimes. The runtimes for the conversion from both the PDB and the SDF format to USMILES are shown in Table 9.

Table 9. Runtimes for the Conversion of the Ligand Expo Data Set from PDB and SDF to USMILES

data set	entries	runtime (s)
PDB (all)	602704	147
SDF (all)		79
PDB (>7 atoms)	204797	110
SDF (>7 atoms)		64

The conversion from SDF provides a point of reference for the performance of our method, since the steps after the generation of the valence bond structure are identical for both formats. Due to the numerous monatomic and small molecules (e.g., solvent molecules) in the data set, we also used a subset where all entries with less than eight atoms have been excluded. This data set provides a more realistic picture of the average runtimes per molecule. The molecule entries in the PDB format were only supplied as single files in a tar archive, which can cause large IO overhead. To avoid this, we concatenated all files into one large file which is a common procedure for other formats such as SDF.

Time measurements were performed on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. The average runtime for the conversion of a single molecule from the PDB format is approximately 1 ms. The comparison to the value obtained for the SDF format (0.4 ms/molecule) shows, that the runtimes lie well in the range of conventional file format conversions. Our method can hence be used even in large scale applications.

■ CONCLUSION

We have presented a novel method for the perception of molecular structures from atomic coordinates. This method is based on the recently published NAOMI model,[11] which has been developed for the appropriate representation of organic molecules. The robustness of our approach has been assessed by processing the Ligand Expo data set in PDB format and comparing the resulting molecules to the structures from the corresponding SDF files. The results are correct in more than

95% of the cases showing that our method is able to produce reasonable results even when working with coordinates of varying quality. The method's accuracy has been demonstrated by comparison to manually curated molecules from previously published benchmarking sets. Our method was successful in 98% of the cases and was able to generate reasonable molecular representations even from structures with distorted geometries. A direct comparison to the tools fconv, I-interpret, and MOE shows that the combination of geometrical and chemical criteria used in our method is the key to avoid many assignment problems. Due to the average runtime of less than 1 ms per molecule the method is perfectly suitable for large scale applications.

Since the method is based on the NAOMI model, it is currently limited to organic molecules which can be represented by valence bond structures. This limitation does, however, only exclude a small number of molecules in the PDB and is thus considered acceptable. Because of missing hydrogen atoms and low resolution of most PDB entries the appropriate tautomeric form can usually not be deduced from the atomic coordinates alone. This would require a more advanced analysis of the ligand's energy or the explicit consideration of the molecule's environment, for example, the binding pocket of the protein, neither of which are in the scope of our method. The method is included in the current version of the NAOMI-converter which can be downloaded at http://www.zbh.uni-hamburg.de/naomi. It is available free of charge for academic use.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

PDB files of the 563 molecules used in the validation studies, the corresponding USMILES of the reference structures, and the converted molecules for which the perception was considered incorrect are provided. 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 Addresses

[§]Georg Simon Ohm University of Applied Sciences, Kesslerplatz 12, 90121 Nuremberg, Germany.
[||]Evotec AG, Essener Bogen 7, 22419 Hamburg.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) PDB File Formats. http://www.pdb.org/pdb/static.do?p=file_formats/index.jsp (accessed Oct 19, 2011).

(3) PDB Format, version 3.3. http://www.wwpdb.org/documentation/format33/v3.3.html (accessed Oct 19, 2011).

(4) Meng, E.; Lewis, R. Determination of Molecular Topology and Atomic Hybridization States from Heavy Atom Coordinates. *J. Comput. Chem.* **1991**, *12*, 891−898.

(5) Baber, J.; Hodgkin, E. Automatic Assignment of Chemical Connectivity to Organic Molecules in the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401−406.

(6) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.

(7) Labute, P. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2005**, *45*, 215−221.

(8) Froeyen, M.; Herdewijn, P. Correct Bond Order Assignment in a Molecular Framework Using Integer Linear Programming with Application to Molecules Where Only Non-Hydrogen Atom Coordinates Are Available. *J. Chem. Inf. Model.* **2005**, *45*, 1267−1274.

(9) Zhao, Y.; Cheng, T.; Wang, R. Automatic Perception of Organic Molecules Based on Essential Structural Information. *J. Chem. Inf. Model.* **2007**, *47*, 1379−1385.

(10) Sayle, R. PDB: Cruft to Content (Perception of Molecular Connectivity from 3D Coordinates). Daylight Chemical Information Systems Inc. MUG'01 Presentation, 2001. http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html (accessed Oct 18, 2011).

(11) Urbaczek, S.; Kolodzik, A.; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(12) Cordero, B.; Gomez, V.; Platero-Prats, A. E.; Reves, M.; Echeverria, J.; Cremades, E.; Barragan, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832−2838.

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

(14) Hartshorn, M.; Verdonk, M.; Chessari, G.; Brewerton, S.; Mooij, W.; Mortenson, P.; Murray, C. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(15) I-Interpret, version 1.0, Shanghai Institute of Organic Chemistry. http://www.sioc-ccbg.ac.cn/?p=42 software=i-interpret (accessed Oct 4, 2012).

(16) fconv—A tool not only for file conversion, version 1.24, Gerd Neudert, University of Marburg. http://pc1664.pharmazie.uni-marburg.de/drugscore/fconv_download.php (accessed Oct 4, 2012).

(17) Molecular Operating Environment (MOE), version 2011.10, Chemical Computing Group Inc. http://www.chemcomp.com/software.htm, (accessed Oct 4, 2012).

(18) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.

(19) Ligand Expo, RCSB PDB. http://ligand-expo.rcsb.org/ (SDF and PDB dataset downloaded Jul 10, 2012).