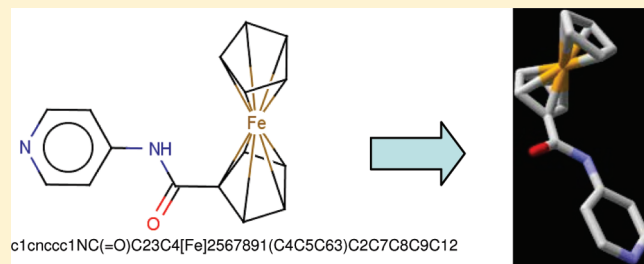


# Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress

Alessio Andronico,<sup>†,§</sup> Arlo Randall,<sup>†,§</sup> Ryan W. Benz,<sup>†,||</sup> and Pierre Baldi<sup>\*,†,‡</sup><sup>†</sup>School of Information and Computer Sciences, Institute for Genomics and Bioinformatics and <sup>‡</sup>Department of Biological Chemistry, University of California, Irvine, Irvine, California 92697-3435, United States

**ABSTRACT:** Accurate prediction of the 3-D structure of small molecules is essential in order to understand their physical, chemical, and biological properties, including how they interact with other molecules. Here, we survey the field of high-throughput methods for 3-D structure prediction and set up new target specifications for the next generation of methods. We then introduce COSMOS, a novel data-driven prediction method that utilizes libraries of fragment and torsion angle parameters. We illustrate COSMOS using parameters extracted from the Cambridge Structural Database (CSD) by analyzing their distribution and then evaluating the system's performance in terms of speed, coverage, and accuracy. Results show that COSMOS represents a significant improvement when compared to state-of-the-art prediction methods, particularly in terms of coverage of complex molecular structures, including metal–organics. COSMOS can predict structures for 96.4% of the molecules in the CSD (99.6% organic, 94.6% metal–organic), whereas the widely used commercial method CORINA predicts structures for 68.5% (98.5% organic, 51.6% metal–organic). On the common subset of molecules predicted by both methods, COSMOS makes predictions with an average speed per molecule of 0.15 s (0.10 s organic, 0.21 s metal–organic) and an average rmsd of 1.57 Å (1.26 Å organic, 1.90 Å metal–organic), and CORINA makes predictions with an average speed per molecule of 0.13 s (0.18 s organic, 0.08 s metal–organic) and an average rmsd of 1.60 Å (1.13 Å organic, 2.11 Å metal–organic). COSMOS is available through the ChemDB chemoinformatics Web portal at <http://cdb.ics.uci.edu/>.



## INTRODUCTION

Accurate prediction of the 3-D structure of small molecules is essential in order to understand their physical, chemical, and biological properties, including how they react or interact with other molecules. The fundamental importance of 3-D structure prediction on a large-scale has increased dramatically with the recent development and free availability of large, public, databases of small molecules<sup>1–4</sup> containing millions of compounds coupled with the growing interest in the design of new molecules<sup>5,6</sup> and the exploration of the astronomically sized space of virtual molecules that are not readily available but could be synthesized from existing building blocks.<sup>7</sup> In contrast, the current version of the Cambridge Structural Database (CSD),<sup>6,9</sup> the main repository of experimentally determined small molecule structures that is available only through commercial licenses, contains less than half a million structures. Thus, the overwhelming majority of small molecules is not represented in the CSD, and there is an important need for methods capable of accurately predicting molecular structures on a high-throughput scale.

The study of computational 3-D structure generation has a long history, going back more than 30 years.<sup>10</sup> Current methods for predicting 3-D structures suffer from problems of speed, coverage, or accuracy. On one hand, quantum mechanics or molecular mechanics methods<sup>11–14</sup> are often capable of yielding

accurate structures. However, these methods often depend on a good starting structure and typically require on the order of minutes to hours to generate a low-energy structure on current computers. So, while these may be the methods of choice for a fine-grained analysis of a small set of molecules of interest, they are not suitable for sifting through very large sets of molecules. On the other hand, approaches based on simple rules or templates are much faster, can be used on a large-scale, and have led in some cases to widely used commercial products.<sup>10,15</sup> However, these methods suffer from issues of coverage or accuracy and fail to produce models for some complex organic molecules. In addition, current programs employing these methods typically fail on the majority of metal–organic molecules in the CSD. This shortcoming is critical because metal–organic molecules are used in a variety of important medical and industrial applications. Furthermore, because the best methods in this class are embedded in commercial products, they are not open or freely available to the academic research community. Thus, there is a clear need for a new system for predicting the 3-D structures of small molecules that is fast, accurate, has good coverage, and is open and free for the academic research community.

Received: June 8, 2010

Published: March 18, 2011

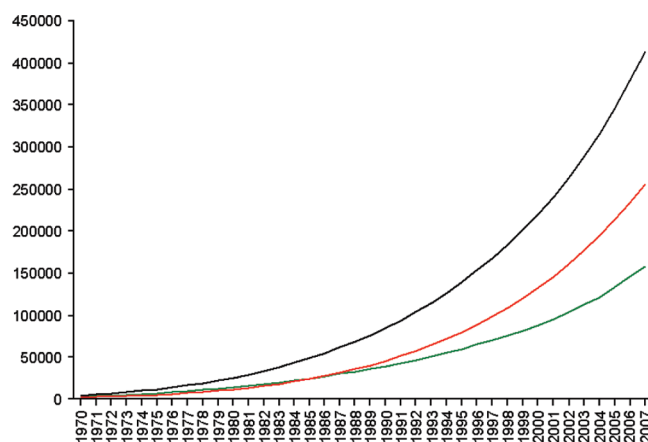
In the somewhat related field of protein structure prediction, successful methods have been developed that rely heavily on the freely available Protein Data Bank (PDB)<sup>16</sup> in one way or another, for instance, by building models using homology to known structures, by including statistical (or knowledge-based) potential terms in the energy function to be minimized, or by extracting libraries of fragments (and their coordinates) from the PDB and using them as building blocks that are then pieced together by some deterministic or stochastic algorithm.<sup>17–19</sup> This suggests the possibility of using a similar approach for small molecule structure prediction by relying on a library of fragment parameters that can be used to reconstruct more complex 3-D structures. Thus, our goal is to present this idea in detail and build the first version of a system capable of predicting 3-D structures rapidly, accurately, with good chemical coverage, that can scale to large databases or combinatorial libraries, and that is as free and open for academic research as possible. Although, we illustrate the general approach using a fragment library extracted from the CSD, it should be clear that the approach can be applied using any fragment library.

In this article, we present the methods and benchmark the performance of a novel data-driven system called COSMOS (COordinate of Small MOlecules) for the high-throughput prediction of small molecule 3-D structures. In essence, the system decomposes any molecule into fragments that are matched to libraries of fragments extracted for illustration purposes from the CSD and assembles the resulting matching fragments into a 3-D model. The presentation follows a broader review perspective that also covers key issues related to small molecule structure prediction. The rest of the article is organized as follows. First, a short summary section on the growth and composition of the CSD is presented. This is followed by a brief review section on the current state-of-the-art in the field and a set of target specifications for the next generation of prediction methods are proposed. Then a section on fundamental methodological issues for the field discusses structural ambiguities and assessment protocols. This is followed by a section that presents our main results obtained by using a fragment library extracted from the CSD, with a focus on the statistical distributions of various fragment types. Then the fragment matching methods and the procedure for building molecules from fragments are described in their own sections. In the final section, the benchmark performance of the new prediction method is assessed using various metrics for coverage, speed, and accuracy.

## ■ GROWTH AND COMPOSITION OF THE CAMBRIDGE STRUCTURAL DATABASE

**Growth of the CSD.** Because here we illustrate our approach using the CSD, we first briefly review several statistics regarding the CSD, starting with its growth. Taking snapshots of the number of entries with error-free 3-D coordinates in the CSD in recent decades demonstrates significant growth: 25,016 structures in 1980, 84,068 structures in 1990, 220,013 structures in 2000, and 425,122 in the 2009 release of the CSD (CSD09) (Figure 1). In fact, the CSD has been growing exponentially, roughly doubling every 7–8 years.

**Composition of the CSD: Importance of Organometallic Molecules.** The growth within the CSD has been increasingly dominated by organometallic (also known as metal–organic) molecules, with their representation growing from 44.9% in 1980 to 53.8% in 1990, 60.1% in 2000, and 61.7% in 2009. The



**Figure 1.** CSD molecules growth over four decades. Green = organics. Red = metal–organics.

increasing numbers of organometallic molecules crystallized in recent years results probably from multiple factors including the difficulty of characterizing their structures by non-X-ray methods and their increasing relevance in several areas of research. Organometallics already play essential roles in medicine, chemical research, and industry. As a broad class, organometallic small molecules are much more complex and structurally diverse than organic small molecules; thus, they can exhibit functional properties that cannot be realized by any organic molecule. Their complexity and diversity also present a significant challenge for structure prediction methods and prediction failure rates are much higher on organometallics. Thus, there is a significant need for new prediction methods that can accurately predict the structures of organometallic molecules.

While organic small molecules consist only of the following atoms: H, B, C, N, O, F, P, S, Cl, Br, and I, organometallics contain combinations of metal atoms and atoms from the organic subset. Organometallics have a much larger set of atomic building blocks and can exhibit much more complex connectivity between atoms than organic molecules. Two simple measures of molecular complexity are the number of atoms and the maximum coordination number (CN) observed among any of its atoms. The average number of heavy atoms in the organic molecules in the CSD09 is 20.0, and the average maximum CN is 3.5. The average number of heavy atoms in organometallics is 34.5, and the average maximum CN is 6.1. Less than 1% of organic molecules have a maximum CN of more than four; in contrast, more than 70% of organometallic molecules have a maximum CN of more than four.

From a significance standpoint, organometallics have important applications in chemical research and industry, for instance, as catalysts in the large scale production of chemicals used in pharmaceuticals, food products, and fuel additives. Organometallic small molecules are also playing an increasing role in medicine. Most drugs are organic small molecules; however, 45 of the 1,382 drugs (3.2%) in the Drug-Bank database<sup>20,21</sup> listed as approved are organometallic, and 145 of the 3,203 drugs (4.5%) listed as experimental are organometallic. Only three of the approved organometallic drugs (0.22%) have a maximum CN of more than four, and two of them are only slightly different versions of vitamin B-12 (cyanocobalamin and hydroxocobalamin). There are 24 experimental organometallic drugs (0.75%) with a maximum CN of more than four. A variety of organometallic

Table 1. Comparison of Four Programs Using a Random Sample of 10,000 Molecules from CSD09<sup>a</sup>

	all (10,000)		organic (3,444)		metal–organic (6,556)	
	time (s)	cov. (%)	time (s)	cov. (%)	time (s)	cov. (%)
DG	14.6	100.0	8.16	100.0	18.0	100.0
FROG	NA	30.8	NA	86.1	NA	1.7
Open Babel	28.4	31.1	27.5	85.7	45.1	2.4
CORINA	0.04 (0.20)	66.4	0.02 (0.05)	98.5	0.07 (0.31)	49.6

<sup>a</sup> Time represents the average prediction time per molecule in seconds, computed over all molecules for which the program outputs a prediction. Times shown in parentheses represent the average prediction times per molecule, including all molecules for which a prediction is attempted. Accurate prediction time for FROG is not available (NA) because it is only available as a Web server. Coverage (cov.) is expressed as a percentage of the total number of molecules in the corresponding class.

molecules have been shown to have anticancer activity.<sup>22,23</sup> Organometallic molecules also act as delivery systems for vital minerals such as iron, magnesium, and zinc.

### ■ CURRENT STATE-OF-THE-ART IN SMALL MOLECULE 3-D STRUCTURE PREDICTION AND SPECIFICATIONS

**State-of-the-Art.** The 3-D structure of small molecules can be predicted fairly accurately using quantum mechanics (QM) or molecular mechanics (MM) approaches.<sup>11–14,24</sup> However, these methods can still produce imperfect models and suffer from two drawbacks: they often require a good starting configuration, and they require a nontrivial computational time. MM may require from seconds to minutes for a small organic molecule and scale as the square of the number of atoms ( $n^2$ ) involved. For metal–organic molecules, in general there are no proper MM force fields, and thus, one must typically resort to QM methods, which may take from minutes to hours and scale as the cube ( $n^3$ ) or higher. In either case, these durations are not compatible with the prediction of the structures of millions or billions of molecules. Furthermore, both QM and MM methods benefit from a good starting point and thus ought to be viewed as refined methods that can be used, in conjunction with the high-throughput methods described here, to produce refined structures for a small number of suitably selected molecules of interest. In particular, QM and MM methods can be used to build libraries of basic fragments with even greater coverage than what can be extracted from the CSD.

For the rapid, high-throughput, prediction of molecular 3-D structures, several methods have been proposed over the years ranging from methods based on distance geometry to rule-based methods.<sup>10,25–30</sup> Some of these methods have been incorporated into widely used commercial products for general coordinate generation including Concord (<http://www.tripos.com><sup>27,31</sup>), CORINA (<http://www.mol-net.de/><sup>10,15</sup>), LigPrep (<http://www.schrodinger.com/products/14/10/>), and Omega (<http://www.eyesopen.com>). Probably the most widely used of these programs is CORINA, which has been licensed in the past to predict 3-D structures for some of the molecules in the main large public databases of small molecules such as PubChem (<http://pubchem.ncbi.nlm.nih.gov/><sup>4,32</sup>) and ChemDB.<sup>2,3</sup> Thus we have obtained a license for CORINA for evaluation and comparison purposes. While commercial methods have demonstrated some success in the field of small molecule structure prediction, there has been a general lack of comparable programs that are freely available to the academic community. In particular, public chemical databases such as PubChem, ZINC, and ChemDB could benefit from a high quality free method. A few free

programs have emerged, such as Open Babel ([http://open-babel.org/wiki/Main\\_Page](http://open-babel.org/wiki/Main_Page)<sup>33</sup>) and FROG (<http://bioserv.rpbs.univ-paris-diderot.fr/cgi-bin/Frog2><sup>30</sup>). However, the performance gap between these free methods and the commercial methods is significant.

**Evaluation.** Here, we evaluate the coverage, speed, and accuracy of some relevant existing methods on a random subset of 10,000 molecules extracted from the 2009 version of the CSD (CSD09). All timing benchmarks in this paper are carried on a single workstation with AMD Opteron Processor 250 (2.4 Gz processor), 4 GB of RAM, and 1 MB of cache. We tested the free methods Open Babel and FROG, the commercial method CORINA (version 3.46), and a distance geometry (DG) approach similar to that described in refs 29 and 34 and developed in-house. The DG method is evaluated alone, without a subsequent energy minimization step, to assess how this general approach performs while retaining its fast conversion speed; adding an energy minimization step significantly increases the DG conversion times. The results of the benchmark tests are summarized in Tables 1 and 2. Table 1 displays the percentage of molecules for which a prediction was returned (coverage) and the average prediction time for the entire set of 10,000 molecules as well as the organic (3,444 molecules) and metal–organic subsets (6,556 molecules). In order to make a fair comparison of the methods, we identified the common subset of molecules for which all methods were able to produce models. The common subset includes 2,443 organic molecules and only 47 metal–organic molecules. Table 2 displays the average rmsd, the percentage of models containing atom–atom clashes, and the average prediction time of the methods on the common subset (for details about the definition of atom clash, see Molecular Fragments, Statistics, and Representations).

**Coverage.** The DG method has perfect coverage of the organic and metal–organic molecules, while CORINA covers 98.5% of organics and approximately half of metal–organics (49.6%). Open Babel covers 85.7% of the organics and only 2.4% of the metal–organics. Similarly, FROG covers 86.1% of the organics and only 1.7% of the metal–organics.

**Speed.** CORINA is the fastest method with an average prediction time of 0.01 s on the common subset of molecules and 0.04 s calculated on all the molecules on which it was able to make a prediction. For some additional molecules, CORINA attempted to make a prediction but then failed. When the run-times on these failed molecules are included, the average time to produce a model is 0.20 s. The DG method had an average prediction time of 0.21 s on the common subset molecules and 14.60 s when calculated on the entire set of 10,000 molecules.



**Table 2.** Comparison of the programs reported in Table 1 using the common subset of molecules for which all the four methods were able to generate 3-D structures<sup>a</sup>

	rmsd (Å)	clash %	time (s)
common set (2,490 molecules)			
DG	1.79	10.0	0.21
FROG	1.32	0.6	NA
Open Babel	1.20	0.8	23.9
CORINA	0.95	2.8	0.01
organic (2,443 molecules)			
DG	1.79	10.2	0.21
FROG	1.31	0.6	NA
Open Babel	1.20	0.9	23.7
CORINA	0.95	2.8	0.01
metal–organic (47 molecules)			
DG	1.88	2.1	0.28
FROG	1.42	4.3	NA
Open Babel	1.34	0.0	34.2
CORINA	1.18	2.1	0.01

<sup>a</sup> Average root mean square deviation (rmsd) per molecule is in Angstroms. The clash percentage column reports the percentage of returned models containing atom clashes. Time represents the average prediction time per molecule in seconds. Accurate prediction time for FROG is not available (NA) because it is only available as a Web server.

Open Babel had an average time of 23.9 s on the common subset and 28.4 s over all the molecules it was able to predict. The most recent version of FROG is only available as a Web server (the authors plan to release a downloadable version in the future); thus, we could not accurately calculate an average run-time. The average prediction time calculated from the server response times was approximately 7 s.

**Accuracy.** CORINA is the most accurate on both the organic and metal–organic subsets with an average rmsd of 0.95 Å and 1.18 Å, respectively. There is a noticeable performance gap between CORINA and the next best method, Open Babel, which has an average of 1.20 Å on the organics and 1.34 Å on the metal–organics. FROG has slightly higher averages than Open Babel on both subsets: 1.31 Å on the organics and 1.42 Å on the metal–organics. DG is significantly worse than these with averages of 1.79 Å on organics and 1.88 Å on the metal–organics. Regarding atom–atom clashes, less than 1% of the models produced by Open Babel and FROG contain clashes. A total of 2.8% of the CORINA models contain clashes, and 10.0% of the DG models contain clashes.

In summary, the commercial method CORINA is the fastest and most accurate method among the ones we tested. In addition, CORINA covers nearly all organic molecules and about half of the metal–organics. This coverage is significantly better than both FROG and Open Babel. The obvious strength of the DG method is that it can make predictions on virtually any molecule; however, the weak accuracy and speed results indicate that it would be a poor choice to use as a stand-alone method. The DG accuracy could be improved by adding an energy minimization step but at the cost of further increasing the run time. On the basis of the benchmark tests, CORINA is clearly one the best existing methods for high-throughput small molecule structure prediction; thus, only CORINA is used for all of

the comparisons with COSMOS presented in the Results section. CORINA's most glaring weakness is that it performs very poorly on metal–organic molecules when compared to organic molecules in terms of speed and accuracy but even more significantly in coverage. One of the key challenges associated with metal–organic molecules is how to handle the molecules with high degrees of connectivity between atoms. For instance, 29% of the metal–organic molecules in the CSD09 have a maximum CN of 7 or higher, and CORINA does not attempt prediction on any molecule with maximum CN greater than 6.

**Challenges Associated with Metal–Organics.** One major difficulty associated with the higher CN atoms is that the 1-D and 2-D molecular representations containing these atoms can correspond to an under-constrained system. The stereo descriptors in these representations can be used to uniquely define the placement of neighboring atoms around an atom with CN = 6; however, this cannot be applied to atoms with higher CN. In some cases, the connectivity among the entire set of atoms defines a unique solution in 3-D space. However, in other cases the system is under-constrained, and many unique 3-D structures will satisfy the constraints provided by the input. In this situation, it is still preferable to produce one or more plausible models than to just fail, as long as the issue is communicated to the user of the predictor. Another challenge associated with predicting metal–organic molecules that is specific to data-driven methods is that, in general, a fragment from a query metal–organic molecule is much less likely to match against a library of fragments than an organic fragment. This results from the fact that the population of metal–organic fragments is much more diverse than the population of organic fragments. The diversity of various types of organic and metal–organic fragments are discussed and analyzed in detail in the section on Molecular Fragments, Statistics, and Representations.

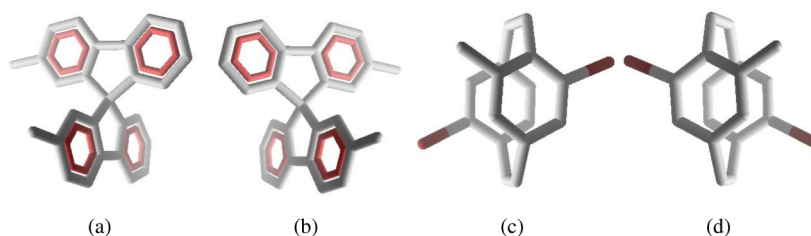
**Specifications.** Given the results discussed above, the current state-of-the-art in small molecule structure prediction falls short of what will be required by the next generation of high-throughput computational and laboratory experiments that rely on predicted structural models. Here we outline new target specifications for the next generation of high-throughput small molecule structure prediction methods. On the basis of current state-of-the-art, for a new prediction method to have an impact in the field, it must excel in one or more (preferably all) of the following areas.

**Coverage.** The system should cover all organic and metal–organic small molecules.

**Speed.** The system should be very fast. It should be suitable for high-throughput applications and capable of processing on the order of one million small molecules per hour on a small cluster (10–100 nodes). Thus, the average prediction time per molecule should be on the order of 0.1 s on a single node.

**Accuracy.** The system should be very accurate. The average root mean square deviation (rmsd), a standard measure of performance in 3-D structure prediction, should not exceed 1.2 Å when computed on a large set of small molecules, consisting of organic and metal–organic molecules. This target accuracy takes into account structural ambiguities discussed in the section on Fundamental Methodological Issues. In addition, the predicted models should have correct coordination geometries and be as free of atom–atom clashes as experimentally determined models.

**Openness.** The system should be freely available for academic research purposes.



**Figure 2.** Rigid fragment structural ambiguities. The rigid fragments in (a) and (b) can be distinguished in their SMILES by including stereochemistry information. (a): Stereo SMILES. Cc1ccc-2c(c1)[C@]3(c4c2cccc4)c5ccccc5-c6c3cc(cc6)C. (b): Stereo SMILES. Cc1ccc-2c(c1)[C@@]3(c4c2cccc4)c5ccccc5-c6c3cc(cc6)C. If the stereochemistry of the central tetrahedral atom is not specified, then the SMILES calculated from the rigid fragments in both (a) and (b) are identical and equal to Cc1ccc-2c(c1)C3(c4c2cccc4)c5ccccc5-c6c3cc(cc6)C. However, the rmsd between the two structures is 3.46 Å. The fully specified SMILES for the rigid fragments in (c) and (d) is Cc1cc2cc(c1CCc3ccc(c3)Br)CC2)Br. The rigid fragment in (c) is the enantiomeric structure of the rigid fragment in (d). The rmsd between the two structures is 2.52 Å.

## FUNDAMENTAL METHODOLOGICAL ISSUES

**General Approach.** The basic approach developed in this work consists of three steps: (1) build fragment libraries, (2) for any query molecule, decompose the molecule into elementary fragments and match those against the libraries, and (3) use the matching fragments to produce a 3-D structure prediction. In turn, there are several possibilities for implementing each step such as (1) the nature and size of the fragments (e.g., rigid fragments) and their origin (e.g., CSD vs other sources), (2) the matching strategy (e.g., exact, partial, fuzzy), and (3) the algorithms for assembling fragments together (e.g., deterministic versus stochastic). For each step, we describe the implementation of the current version of COSMOS. Before addressing these technical issues, we briefly discuss molecular representations, structural ambiguities, and assessment methodology.

**Inputs and Outputs: SMILES and SDF Representations.** Molecules can be represented in many computer formats, such as Simplified Molecular Input Line Entry Specification (SMILES) strings, connectivity matrices (e.g., sdf files), and fingerprints.<sup>35,36</sup> Programs for converting from one format to another exist; thus, in principle, one can work with any format. SMILES is a chemical language that can be used to specify the connectivity and stereochemistry of a molecule in a straightforward and compact line of text.<sup>37–39</sup> The language was extended and improved by DAYLIGHT Chemical Information Systems, Inc. (<http://www.daylight.com>). In SMILES, branches are represented by parentheses, and ring closures are indicated by matching integer tags. Tetrahedral and bond stereochemistry can also be defined with SMILES. The SMILES format is supported by most chemoinformatics software tools and is used extensively to store the 2-D structure of molecules in various public databases. For instance, PubChem,<sup>4,32,40</sup> ChemDB,<sup>2,3</sup> ZINC,<sup>1</sup> and DrugBank<sup>21</sup> all utilize SMILES. Internally, COSMOS takes SMILES input and produces 3-D structure output using the standard sdf (structure-data file) format. However, because these formats can be converted to the other standard representations, users of the system can provide any standard input format and receive any standard output format.

**Structural Ambiguities.** There are several sources of ambiguity between 1-D or 2-D molecular representations and 3-D structure that must be considered and accounted for during the development of a prediction system. The SMILES string or the bond matrix/graph of a molecule does not always specify a unique 3-D structure. Ambiguities can arise even within a single rigid fragment due to the presence of mirror symmetries

(enantiomers) or tetrahedral atoms with unspecified stereochemistry (stereoisomers). These ambiguities can combine exponentially when a molecule with multiple rigid fragments is considered. Some degree of stereochemistry can be specified in the SMILES. However, this does not always occur, and even when it occurs, the specification may be insufficient to describe a unique structure. These ambiguity problems are common in current large databases of molecules and their associated SMILES representations.

Some of these issues are illustrated in Figure 2 using simple rigid fragments. Panels (a) and (b) of Figure 2 display fragments with two different tetrahedral arrangements. The tetrahedral variation can be explicitly indicated in the corresponding SMILES, in which case the correct unique structure ought to be produced. But if the tetrahedral annotation is omitted from the SMILES, then the SMILES does not specify a unique structure. If not specified, then it can be modeled in two possible ways. In some cases one of the options is clearly superior energetically, and thus, the ambiguity is not a problem in model building. However, if both options are energetically equal (as in Figure 2a, b) then either option is equally likely. A simple solution is to notify the user of the source of ambiguity and to return both models. However, these issues can multiply with the number of tetrahedral atoms within a rigid fragment and with the number of fragments in a molecule. Approximately 28% of the molecules in CSD09 have one or more undefined stereo centers in the SMILES representation when the source SDF file properties alone are used to generate the SMILES.

There are other well-known sources of ambiguity that cannot be incorporated in the current SMILES notation and result in a SMILES string being associated with multiple 3-D structures with identical or similar energies. Clear examples of this phenomenon are provided by mirror images, as shown in panels (c) and (d) of Figure 2. In the rigid fragments from CSD09, we find that 95% of the SMILES of the original and the mirror image are identical. For the vast majority of these, the original and the mirror have the same 3-D structure due to additional symmetries; however, the rmsd between the original and the mirror is greater than 0.30 Å for approximately 10% of these. In these cases, the mirror structure is the enantiomer of the original. Approximately 25% of the molecules in CSD09 have one or more rigid fragments with possible enantiomers. Panels (c) and (d) of Figure 2 provide examples of the enantiomeric structures for a single SMILES, and the rmsd between the structures is 2.52 Å. The solution to this source of ambiguity is to check the SMILES of the mirror image of the rigid fragment being considered. If it

matches the original, then the rmsd between the model and its mirror image can be checked. If the rmsd is significant, then both models should be returned, and the user should be notified of the reason. Again, however, this can become impractical because of the combinatorial explosion across multiple rigid fragments. Approximately 50% of the SMILES from the molecules in CSD09 have at least one rigid fragment flagged for ambiguity because of undefined stereo centers or enantiomers.

At the level of whole molecules, rotatable bonds introduce another source of ambiguity that must be dealt with appropriately. The experimentally determined coordinates of a small molecule in the CSD represents one low-energy conformation, but when the molecule has rotatable bonds it may have many conformations that are approximately equal energetically.

These sources of ambiguity can be handled in a variety of ways both in how a prediction method is designed and during assessment. In cases of ambiguity, returning no prediction is a poor solution. In many applications (e.g., docking<sup>41</sup>), having one low energy structure is better than having none. Likewise, QM or MM methods cannot be used to perform further refinements, when no structure is produced. Furthermore, not producing a structure prevents comparison to other predictors that output a structure. The solution we adopt is to (1) always produce at least one structure, sampled from the set of lowest energy structures described by the same SMILES string, (2) alert the user to the presence of ambiguity, (3) produce also a second structure when the ambiguity is only between two alternatives, and (4) when the source of ambiguity is undefined stereocenters, return all stereoisomers (e.g., up to 64 models). Likewise, these sources of ambiguity can be handled differently during assessment. It is possible, for instance, to perform a restricted form of assessment on the rigid fragments only or on the ring systems only. This at least removes a major source of ambiguity (rotatable bonds). Another approach is to remove molecules with ambiguity from the assessment entirely or, if the predictors produce multiple predictions, to take the predictions with the lowest rmsd to the target. All these assessment approaches have their advantages and drawbacks; therefore, we have utilized multiple assessment measures in this work. In the end, however, it is common practice to report primarily the rmsd between the first model returned by the predictor and the experimentally solved structure; thus, this measure is always included in this work.

**Assessment Methodology.** *Accounting for Redundancy.* Speed, accuracy, and coverage of prediction methods can be assessed using the molecules in the CSD. However, great care must be exercised to avoid the circularity of using the entire CSD data to predict CSD structures. This is a standard statistical problem that often comes up in machine-learning and bioinformatics applications. To address this issue, one must run validation experiments where a subset of the data is used to make blind predictions on a different set of the data. In particular, one can use the molecules contained in one older version of the CSD to predict the new structures that have been deposited in a more recent version. One can further reduce the chance of any circularity by removing any residual redundancy between the training and validation sets, as is routinely done in protein structure prediction. This redundancy—reduction is achieved by retaining in the validation set only molecules that have low similarity to the molecules in the training set. In protein structure prediction, it is typical to retain for instance only proteins that

have less than 25% sequence similarity to the sequences in the training set.

As a conservative approach, we used data from the 2004 version of the CSD (CSD04) to build the fragment libraries for predicting the structures of new molecules found in the 2009 version of the CSD (CSD09). This is quite stringent since the CSD almost doubled in size from 2004 to 2009. Furthermore, in the Results section, we present the speed and accuracy results obtained by using a set of 10,000 molecules and also a reduced set of molecules that have low similarity with the molecules in our training set.

Similarity is assessed here by comparing chemical fingerprints<sup>42,43</sup> with the most widely used measure of molecular similarity, the Tanimoto coefficient,<sup>36,44</sup> which scales from 0 to 1. For our reduced set, we retained only the molecules with Tanimoto coefficient below 0.6.

**Accuracy Measures.** The average root-mean-square deviation (rmsd) is the primary measure used to assess the accuracy of predicted models in this work. When a molecule contains topologically equivalent atoms, the rmsd is calculated for all possible mappings between the model and the native structure, and the minimum resulting value is used for assessment. While the rmsd is a good measure for assessing the overall accuracy of predicted model coordinates, it is not appropriate for every type of evaluation (see Structural Ambiguities section). Specifically, an interesting accuracy measure for metal–organic molecules is the percentage of correctly predicted metal coordination geometries. Thus, we have developed a simple angle-based measure for this purpose.

For each CN, 3 through 8, there are two or more metal coordination geometries that occur frequently in the CSD. The specific geometries analyzed in this work are (CN = 3) t-Shaped, trigonal pyramidal, and trigonal planar; (CN = 4) square planar and tetrahedral; (CN = 5) square pyramidal and trigonal bipyramidal; (CN = 6) octahedral and trigonal prismatic; (CN = 7) capped octahedron, capped trigonal prismatic, and pentagonal bipyramidal; and (CN = 8) cubic, dodecahedral, and square antiprismatic.<sup>45</sup> To develop an approach for classifying the coordination geometries, we started by generating an idealized monocentric (central atom and its covalently bonded neighbors only) template structure for each of these geometries. Next, we developed an angle-based similarity metric, where each monocentric fragment is represented as a set of angles ( $M\theta$ ). A given angle in the set,  $M\theta_{i,j}$ , is the angle that neighbor atoms  $i$  and  $j$  form with the central atom, where the central atom is the vertex. The set is composed of the angles calculated using all  $CN(CN - 1)/2$  possible pairs of neighbor atoms. The difference between the model and template angle sets is summarized by the mean difference between their corresponding angles:  $\Delta\theta(M\theta, T\theta) = [\sum_i^{CN} \sum_{j < i}^{CN} |M\theta_{i,j} - T\theta_{i,j}|] / [CN(CN - 1)/2]$ . For the metric to assess the geometry only and ignore the relative placement of atoms, all of the neighbor atoms are considered equivalent.  $\Delta\theta$  is calculated for all possible atom mappings between the model and template fragment, and only the minimum value is kept. Finally, in order to discover an appropriate threshold for classification, we applied the mean difference metric ( $\Delta\theta$ ) to each pair of template geometries with the same CN. The results of this analysis are summarized in Table 3. On the basis of these results, we selected  $\Delta\theta < 10.0^\circ$  as an upper threshold for classifying a model as a specific geometry because it provides clear separation between most of the geometries but also allows for some variability.

To classify the geometry of a fragment, we calculate  $\Delta\theta$  versus all templates with the same CN and identify the best match. If for



the best match  $\overline{\Delta\theta} < 10.0^\circ$ , then the fragment is classified as that geometry, and if not, then it is considered unclassified. If a CSD fragment can be classified, then the corresponding model fragment is considered correct if it has the same geometry class. If a CSD fragment is unclassified, then the ideal templates are ignored, and the corresponding model fragment is compared directly to the CSD fragment. The model fragment is considered correct if  $\overline{\Delta\theta}(M\theta, T\theta) < 10.0^\circ$ , where  $T\theta$  is the angle set of the CSD fragment.

## ■ MOLECULAR FRAGMENTS, STATISTICS, AND REPRESENTATIONS

**Fragment Libraries, Statistics, and Representations.** Molecular modeling methods utilize many kinds of fragments at varying levels of granularity, from entire molecules at the coarser level to single pairs of bonded atoms (e.g., C–C) at the finer level. In this work, primarily two types of intermediate fragment representations were utilized: rigid fragments and cyclic

Table 3. Comparison of the Metal Geometry Templates Using the Mean Angle Difference ( $\overline{\Delta\theta}$ ) Assessment Measure

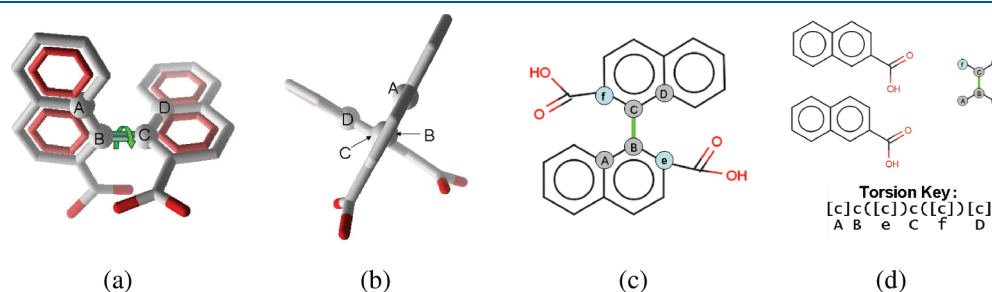
	mean angle difference ( $\Delta\theta$ )
coordination number 3	
t-shaped vs trigonal pyramidal	35.7°
t-shaped vs trigonal planar	40.0°
trigonal pyramidal vs trigonal planar	13.0°
coordination number 4	
square planar vs tetrahedral	36.5°
coordination number 5	
square pyramidal vs trigonal bipyramidal	12.0°
coordination number 6	
octahedral vs trigonal prismatic	22.5°
coordination number 7	
capped octahedron vs capped trigonal prismatic	21.6°
capped octahedron vs pentagonal bipyramidal	14.6°
capped trigonal prismatic vs pentagonal bipyramidal	14.4°
coordination number 8	
cubic vs dodecahedral	19.0°
cubic vs square antiprismatic	15.6°
dodecahedral vs square antiprismatic	8.0°

fragments. These representations are described in detail below. In the determination of fragments, all bonds present in the source files were utilized, including metal pi bonds. In addition, polymeric structures were handled by considering only one copy of the monomer.

*Rigid Fragments and Their Torsion Angles.* Rigid fragments are molecular substructures obtained by fragmenting a molecule at each of its rotatable bonds, such that the resulting fragments contain no rotatable bonds themselves (Figure 3). For each rotatable bond, the torsion angle and the SMILES for the atoms used to calculate the angle, and their first bonded neighbors (“torsion key”) are saved along with the rigid fragments themselves. For a molecule with  $n$  rotatable bonds,  $n + 1$  rigid fragments and  $n$  torsion angles are saved. The molecules in CSD09 fragmented into nearly 2.3 million total rigid fragments, but only 294,856 are unique according to their isomeric SMILES. The 10 most frequently occurring organic and metal–organic rigid fragments are presented in Table 4. How the frequencies are distributed among the unique rigid fragments has important implications for the data-driven prediction approach. In fact, the frequency of occurrence of small molecule fragments have been shown to distribute via power laws.<sup>46</sup> Figure 4 presents a plot of the CSD fragment frequencies on a log–log scale, which clearly illustrates the power law phenomenon. The presence of power laws in the fragment distributions indicates that a relatively small fraction of the rigid fragments occur the vast majority of the time. Although the power law distribution applies to both organic and metal–organic fragments, there is a significant difference in how the organic fragments distribute compared to the metal–organic (Figure 4b). The top 0.1% most frequently occurring organic rigid fragments occur so often that they account for 81.3% of all occurrences. In contrast, the top 10% most common metal–organic rigid fragments account for only 69.3% of all occurrences. The basic implication of this result for small molecule structure prediction is that the organic rigid fragments of new molecules are much more likely to have a match in the data libraries than the metal–organic rigid fragments.





















In this work, a single prototype structure is used to represent the class of rigid fragment structures associated with the same SMILES. The prototype is selected by calculating the rmsd between all pairs of structures in the group and keeping the one with the minimum sum of rmsd values versus all other pairs. For the CSD09, this process yields 294,856 prototype rigid fragments structures.

During the fragmentation of the experimental 3-D structures, the torsion angles between all connected pairs of rigid fragments are calculated and saved along with SMILES representing the



**Figure 3.** Fragmentation at rotatable bonds. Small molecule structures can be divided into independent fragments at each rotatable bond. (a): The rotatable bond is highlighted, and the four atoms used to define the torsion angle are labeled with capital letters. (b): The same molecule rotated 90° degrees on the vertical axis. (c): Schematic representation with additional atoms bound to the torsion bond atoms labeled with lower-case letters. These atoms, in addition to the torsion atoms, are used to define the torsion key SMILES. (d): Resulting fragments and torsion key SMILES. The original SMILES string for the molecules is c1ccc2c(c1)ccc(c2c3c4ccccc4ccc3C(=O)O)C(=O)O.

Table 4. Ten Most Frequently Occurring Organic and Metal-Organic Rigid Fragments<sup>a</sup>

Top Ten Organic Rigid Fragments			Top Ten Metal-Organic Rigid Fragments		
SMILES	2D	% of Total	SMILES	2D	% of Total
c1ccc(cc1)P		7.13%	C[Si](C)(C)C		0.83%
CCC		6.23%	C[Si](C)(C)N		0.49%
COC		5.00%	CC(C)(C)[SiH3]		0.37%
Cc1ccccc1		4.31%	c1ccc(cc1)[SiH3]		0.36%
CC(=O)O		2.92%	CP(C)(C)[Pt]		0.23%
CCO		2.73%	CO[Cu]		0.22%
CC(C)(C)C		2.55%	c1ccc(cc1)[SnH]		0.22%
CCN		2.54%	CC(C)[SiH3]		0.20%
CC(C)C		1.90%	CC[SnH]		0.19%
CNC		1.35%	CP(C)(C)[Ru]		0.19%

<sup>a</sup> The percentage of total calculation considers all organic and metal-organic rigid fragments, and the total number of rigid fragments is 2,273,659.

atoms in the rotor bonds and their first bonded neighbors (Figure 3d and Figure 5). Panel e of Figure 4 shows a plot of the frequency distribution of torsion keys of CSD09 on a log–log scale. The torsion keys clearly distribute via power law, similar to what was observed for the rigid segments. In CSD09, there are approximately 1.9 million torsion angles with 218,391 unique “torsion keys”. To use this torsion information in COSMOS, we first bin all the torsions for a given torsion key in 10° bins. As an example, Figure 5 displays a schematic diagram of the torsion key [cH]c([s])c([cH])[s] and the resulting histogram of torsion angles. We then proceed to save the single value corresponding to the center of the bin with the highest count in the torsion key library. For the torsion key [cH]c([s])c([cH])[s], this single value is 180°. One advantage of this simple univariate approach is that it can be applied consistently regardless of the amount of data available for a given torsion key; however, future versions of COSMOS may utilize multivariate models when appropriate. It has been demonstrated that with sufficient data multivariate distribution models of adjacent torsions can be used to improve the accuracy of model conformations.<sup>47</sup>

**Cyclic Fragments.** The cyclic fragments are extracted from the rigid fragments by removing all non-ring atoms (Figure 6). This yields complete ring systems that can be further broken into individual rings. The complete ring systems and the individual rings are used to build the library.

In CSD09, there are 966,826 complete ring systems, of which 671,097 (69.4%) consist of a single ring. These yield 163,312 unique complete ring systems, including 9,276 unique single rings. The statistical distribution of the ring systems, as shown in panel (c) of Figure 4, also follows a power law, with clear differences between organics and organometallics (Figure 4d). For instance, the top 0.1% most common organic complete ring systems account for 83.5% of all occurrences, and benzene alone accounts for 62.5% of all occurrences. In contrast, the top 10% most common metal–organic complete ring systems account for only 45.0% of all occurrences. The prototype cyclic fragments used to populate the library are selected with the same procedure described above for the rigid fragments.

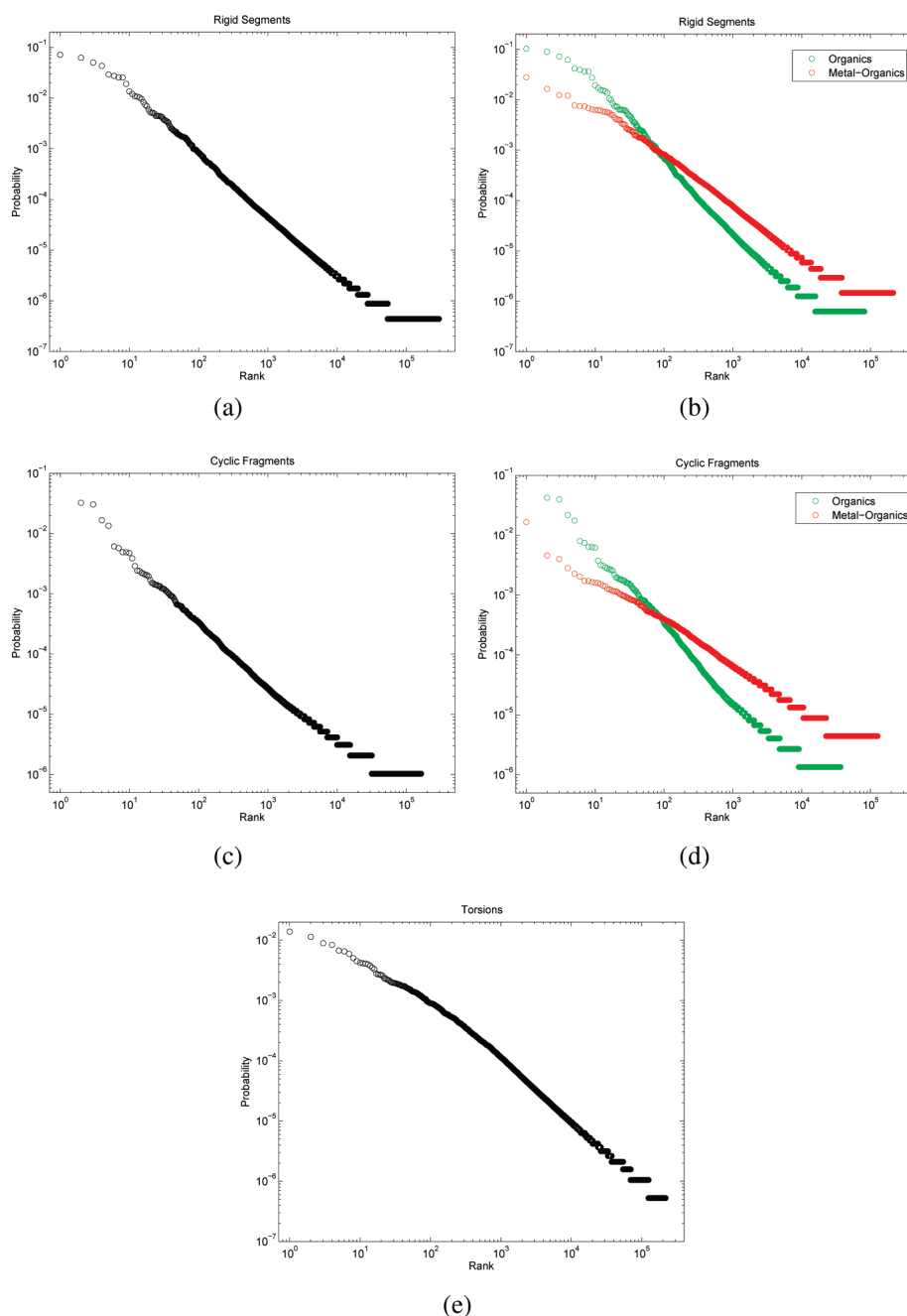
**Atom Clashes.** Here, we analyze the degree of overlap between nonbonded atoms in the molecules in CSD09. To assign a single overlap score ( $\mathcal{Q}$ ) to each structure, we use the following definition:  $\mathcal{Q} = \min_{(ij)} d(a_i a_j) / (v_i + v_j)$ , where  $v_i$  and  $v_j$  are the van der Waals radii<sup>48</sup> of atom  $a_i$  and  $a_j$ , respectively,  $d(a_i a_j)$  is the distance between the two atoms, and the minimum is computed over all pairs of heavy atoms in the molecule separated by more than two bonds. In other words,  $\mathcal{Q}$  quantifies the single most offensive steric clash in a given structure. This measure is also referred to as “overlap factor” and “close contact ratio” and has been used previously in the assessment of crystal structures and predicted models.<sup>15,27,31,49</sup>

In Figure 7, we report the cumulative percentage of molecules in the CSD09 versus the overlap factor ( $\mathcal{Q}$ ). From Figure 7, it is clear that the vast majority of the molecules in CSD09 have an overlap factor greater than 0.7. In fact, only 0.7% of the CSD molecules have an overlap factor below 0.7. As a result of this analysis,  $\mathcal{Q} < 0.7$  was selected as the criteria to categorize a predicted model as containing clashes for assessment purposes.

## ■ MATCHING FRAGMENTS: EXACT MATCHING, PARTIAL MATCHING, FUZZY MATCHING

Given a query molecule, it is decomposed into fragments, and the fragments are matched against the fragment libraries. The matching procedure can be exact when an exact match is found, partial when only a substructure of the fragment is matched, or fuzzy,<sup>50</sup> which allows also for variability in the atom types. One tool for implementing fuzzy matching is provided by the SMARTS language (Daylight), which allows one essentially to create fuzzy SMILES strings. An example of fuzzy matching using SMARTS is given in Figure 8. In many cases, cyclic fragments with different atom types are at one or two positions, but with the same connectivity, have nearly identical 3-D structures. For instance, the basic structure of the porphyrin ring system occurs in a variety of molecules, with different atoms at the center of the structure. Figure 8 shows an example of a SMARTS pattern generated from an unmatched complete ring system (porphyrin portion of chlorophyll A) that could be used to match a complete





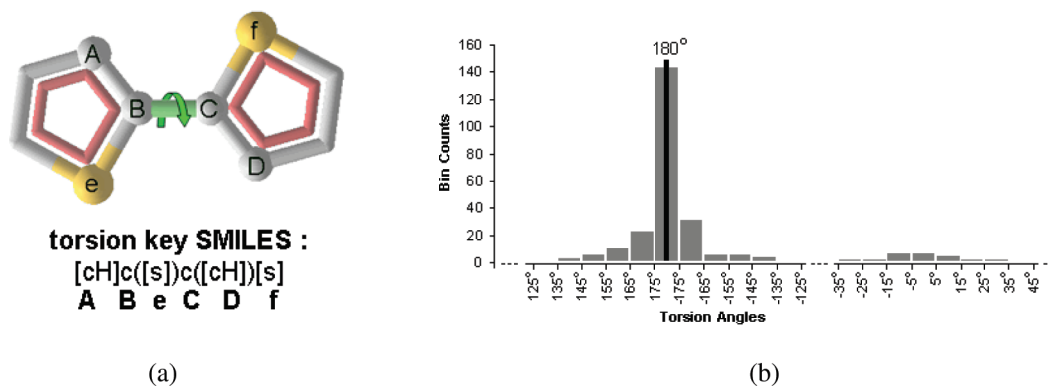
**Figure 4.** Distributions of rigid segments, cyclic fragments, and torsion keys on a log–log scale. The distributions were derived by identifying the unique SMILES for rigid segments, cyclic fragments, and torsion keys, then counting the number of occurrences of each. The probability of each item is plotted versus its frequency rank, such that a rank of 1 corresponds to the most frequently occurring item. The calculated power law distribution exponents are indicated by  $\alpha$ . (a): all rigid segments,  $\alpha = 1.8$ . (b): organic rigid segments,  $\alpha = 1.7$ ; organometallic rigid segments,  $\alpha = 1.9$ . (c): all cyclic fragments,  $\alpha = 1.9$ . (d): organic cyclic fragments,  $\alpha = 1.7$ ; organometallic cyclic fragments,  $\alpha = 2.3$ . (e): torsion keys,  $\alpha = 1.9$ .

ring system analog in the dictionary (porphyrin portion of Heme B). The SMARTS language supports flexibility in the definitions of patterns beyond the simple wildcards shown in Figure 8. It contains a variety of built-in pattern matching, for instance [a] matches any aromatic atom, [#5] matches any nitrogen (n or N), and [#6] matches any carbon atom (C or c). In addition, SMARTS allows for the creation of user-defined patterns. COSMOS uses the following definitions for categorical fuzzy matching: [AM] alkali metals, [AE] alkaline earth metals, [ML] metalloids, [HL] halogens, [NM] non-

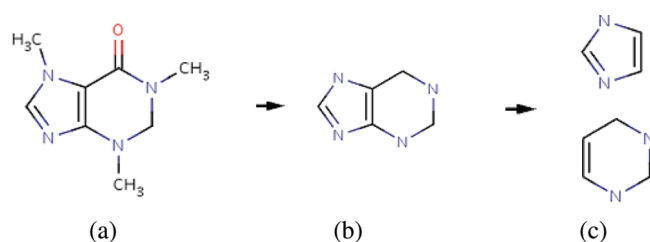
metals, [NG] noble gases, [MT] metals, [TM] transition metals, and [AR] aromatics.

### ■ 3-D STRUCTURE PREDICTION: BUILDING MODELS FROM FRAGMENTS

To a first approximation, there are two general strategies for constructing 3-D models from fragments: (1) a general top-down hierarchical approach, where first large fragments are matched to the libraries as much as possible, followed by smaller



**Figure 5.** Torsion angle bins for the torsion key SMILES: [cH]c([s])c([cH])[s]. This torsion key SMILES occurs 485 times in CSD09. (a): Schematic and SMILES representations of torsion key. (b): Counts of the observed torsion angles in 10° bins. The center of the most populated bin, from 175° to −175°, is 180°.

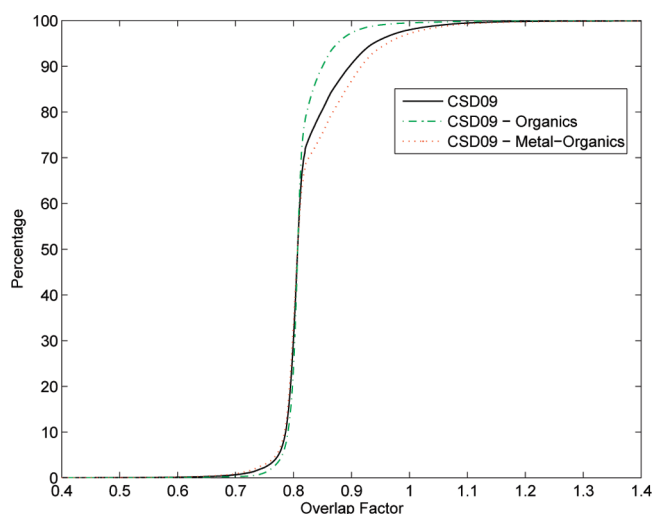


**Figure 6.** Cyclic fragmentation. (a): Starting rigid fragment (caffeine). (b): Terminal atoms are removed from the molecule (caffeine) leaving only the complete ring system (purine). (c): Subrings are extracted from the complete ring system (imidazole and pyrimidine). Using this procedure, three unique cyclic structures (purine, imidazole, and pyrimidine) can be added to the cyclic fragment library.

fragments, or (2) a general bottom-up approach, where only small fragments are matched to the libraries, and then the matches are used to reconstruct the molecules. Here the hierarchical top-down approach used by the current version of COSMOS is described together with the corresponding pseudocode.

Algorithm 1.1 describes the method at the highest level. COSMOS takes a query molecule and extracts its rigid fragments and torsion keys at the rotatable bonds, builds a model for each rigid fragment, reconnects the rigid fragments, and sets the torsion angles of the final model. Building models for the rigid fragments is the most fundamental step, and this process is presented in detail in Algorithm 1.2. COSMOS first checks the fragment library for an exact match. If this fails, then the next step depends on whether or not the rigid fragment contains a ring system.

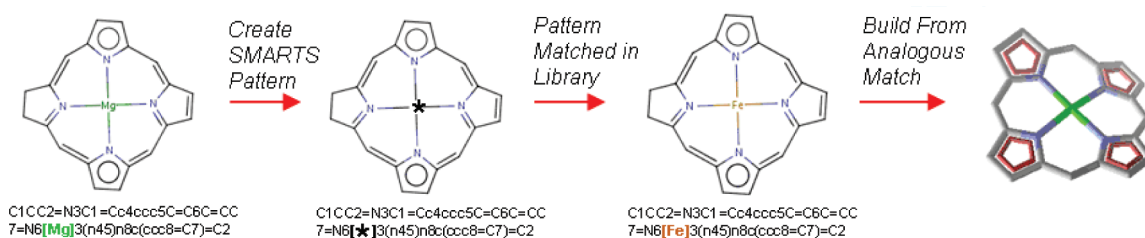
The rigid fragments that contain ring systems are handled by removing all non-ring atoms and using the remaining ring system to search the fragment library using progressively looser matching criteria until a match is found. If the entire ring system cannot be matched in any manner, then the ring system is broken down into individual rings, and they are matched against the fragment library (Algorithm 1.2 is applied to each ring). When an individual ring cannot be matched in the library, GEOMETRIC RING BUILD is called, which constructs the ring in a bottom-up fashion using angles determined by the number of atoms in the ring, their individual properties, and ideal bond lengths.<sup>8,51</sup> Once all of the individual rings have been modeled, they are



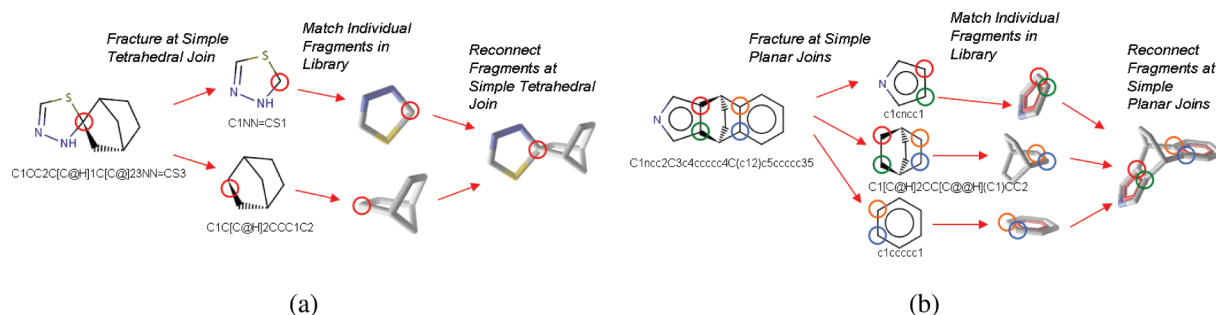
**Figure 7.** Cumulative percentage of molecules in CSD09 vs the overlap factors. Black, green, and red curves represent the entire CSD, the subset of organic molecules, and the subset of metal-organic molecules, respectively.

reconnected to produce coordinates for the entire ring system. Figure 9 provides two examples of how rings can be stitched together. The first example (Figure 9a) corresponds to a simple tetrahedral join. The second example (Figure 9b) corresponds to a simple planar join. The non-ring atoms, which are covalently bonded to atoms in the ring system, are placed using PLACE TERMINAL ATOMS. Each terminal atom is placed using the properties of its source cyclic fragment atom, including stereochemistry, to determine the angle for placement and ideal bond lengths.<sup>8,51</sup>

The rigid fragments, which contain no rings (branched), are handled by breaking them down into monocentric fragments. Each monocentric fragment is matched against the fragment library, and those which are unmatched are built by GEOMETRIC MONOCENTRIC BUILD. This function selects a specific monocentric geometry<sup>45</sup> on the basis of the properties of the central atom, and if there is no way to discriminate a single geometry, then the most prevalent one is chosen. The monocentric fragment is constructed using the selected ideal monocentric geometry and ideal bond lengths. Once all of the individual monocentric fragments have been modeled, they are reconnected by



**Figure 8.** Simple fuzzy matching. Matching the porphyrin ring system of chlorophyll A with the analogous porphyrin ring system of Heme B using the SMARTS pattern language. The asterisk represents a wildcard in SMARTS, so it will match any atom at that position.



**Figure 9.** Fragmenting and reconnecting subfragments of cyclic fragments. The process of breaking down a cyclic fragment into subfragments at simple connections points, matching the subfragments in the library, and reconnecting the fragments, is shown for two types of connection points. (a): Simple tetrahedral joins occur when a single atom is the only connection point between two subfragments of a rigid fragment. These connection points are assembled according to ideal tetrahedral geometry of the connecting tetrahedral atom (circled in red). (b): Simple planar joins occur when two atoms join subfragments, and these atoms plus all of their first bonded neighbors are in the same plane. Planar connections are assembled according to ideal planar geometry. The example shows a molecule with two planar joins, the first with connection atoms circled in red and green, and the second with connection atoms circled in orange and blue.

aligning the overlapping bond vectors, and the dihedral angles are adjusted to minimize steric hindrance.

When all of the rigid fragments of a molecule have been constructed, they are joined together at their rotor bonds by aligning the overlapping bond vectors. Then, each torsion key is matched against the torsion library, and if a match is found, the resulting torsion value is set. If no match is found, then a torsion optimizing procedure built into OEChem is used to set the torsion angle.<sup>52</sup> Considering the 6,795 molecule test set from the Results section, no match was found for 10.5% of the torsion keys, and 23.5% of the molecules had one or more unmatched torsion keys.

## RESULTS

We have described a general framework for developing data-driven predictors for the 3-D structures of small molecules. Here, we describe results obtained with the current implementation where (1) we use only a top-down approach, (2) fragment libraries are composed of rigid fragments and rings, extracted from the old CSD04, and (3) only one prototype per fragment and a single point estimate for each torsion key value are used.

To evaluate COSMOS, we first selected molecules that were unique to CSD09 according to their SMILES. From this subset, we took a random sample of 10,000 molecules to use as a test set (CSD09Test). The set CSD09Test contains 3,597 organic and 6,403 organometallic molecules. Note that CSD09Test is a subset of all the molecules added to CSD after 2004, and that it does not contain any of the molecules that were used to build the fragment libraries.

Coverage, the percentage of molecules predicted among a large set, is the most straightforward assessment of a predictor's ability to generalize. The coverage results of COSMOS and CORINA are summarized in Table 5. COSMOS was able to return a prediction for 96.4% of the molecules with nearly perfect coverage (99.6%) of the organic subset. CORINA performed comparably on organic molecules, but predicted only 51.6% of the organometallic molecules. Of the molecules for which CORINA was not able to produce models, approximately half were skipped because of the presence of atoms with CN > 6, while the other half failed because of various errors that occurred during the attempted prediction. These results indicate that COSMOS represents a significant improvement in coverage of complex molecules.

To objectively assess COSMOS and CORINA, we calculated the accuracy and speed using the subset of molecules in CSD09Test that for which both methods produced predictions. This subset consists of 6,795 molecules and contains 52.1% organic molecules. The predicted structures of each method were compared in terms of (1) rmsd values between the generated structures and their corresponding experimental CSD counterparts, (2) percentage of predicted models that contain clashes, and (3) prediction time. The results are summarized in Table 6. When the entire set is considered, the average accuracy and average speed of the two methods are quite similar. On the organic subset, CORINA is slightly more accurate. It returns fewer models that contain clashes, but it is slower than COSMOS. On the organometallic subset, COSMOS is more accurate and produces fewer models containing atom clashes, but it is slower than CORINA.

In addition to the average accuracy and speed, Table 5 also provides the percentage of molecules with rmsd (1) lower than



**Algorithm 1.1:** COSMOS(*Molecule*, *\$FragmentLibrary*, *\$TorsionLibrary*)

```

Molecule3D ← empty molecule object
for each RigidFragment in Molecule
{
  RigidFragment3D ← BUILDRIGIDFRAGMENT(RigidFragment)
  Connect RigidFragment3D to Molecule3D
}
for each RotatableBond in Molecule
{
  torsion_key ← SMILES of RotatableBond torsion key
  if torsion_key is in $TorsionLibrary {Set RotatableBond torsion angle to $TorsionLibrary[torsion_key]
  else {Set RotatableBond torsion angle using OEChem52
}
return (Molecule3D)

```

**Algorithm 1.2:** BUILDRIGIDFRAGMENTS(*Fragment*, *\$FragmentLibrary*)

```

Fragment3D ← empty molecule object
fragment_key ← SMILES of Fragment
if fragment_key is in $FragmentLibrary {Fragment3D ← $Library[fragment_key]
else if Fragment contains rings
{
  RingSystem ← ring system of Fragment
  ring_system_key ← SMILES of RingSystem
  metals_fuzzy_key ← SMARTS of ring_system_key with metal atoms replaced by categories
  all_fuzzy_key ← SMARTS of ring_system_key with all atoms replaced by categories
  metals_wild_key ← SMARTS of all_fuzzy_key with metal atoms replaced by *
  if ring_system_key is in $FragmentLibrary {Fragment3D ← $Library[fragment_key]
  else if metals_fuzzy_key is in $FragmentLibrary {Fragment3D ← $Library[metals_fuzzy_key]
  else if all_fuzzy_key is in $FragmentLibrary {Fragment3D ← $Library[all_fuzzy_key]
  else if metals_wild_key is in $FragmentLibrary {Fragment3D ← $Library[metals_wild_key]
  else if RingSystem contains multiple rings
  {
    for each IndividualRing in RingSystem
    {
      Ring3D ← BUILDRIGIDFRAGMENT(IndividualRing)
      Connect Ring3D to Fragment3D
    }
  else {Fragment3D ← GEOMETRICRINGBUILD(RingSystem)
  }
  PLACETERMINALATOMS(Fragment3D)
}
else if Fragment does not contain rings
{
  for each MonocentricFragment in Fragment
  {
    mono_fragment_key ← SMILES of MonocentricFragment
    if mono_fragment_key is in $FragmentLibrary {MonocentricFragment3D ← $Library[mono_fragment_key]
    else {MonocentricFragment3D ← GEOMETRICMONOCENTRICBUILD(MonocentricFragment)
    }
    Connect MonocentricFragment3D to Fragment3D
  }
}
return (Fragment3D)

```

**Table 5. Coverage of COSMOS Compared to CORINA<sup>a</sup>**

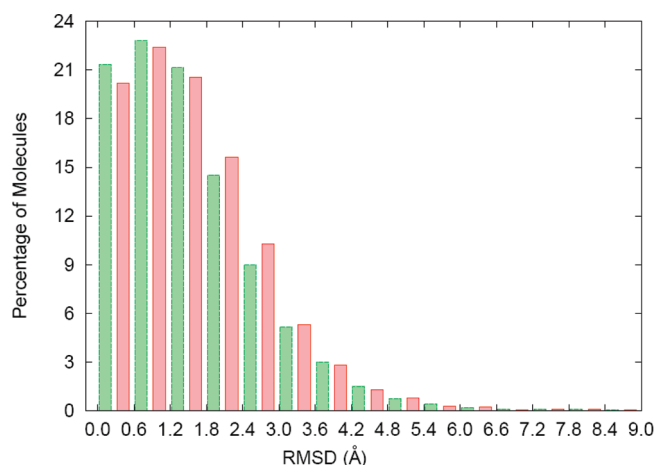
	COSMOS	CORINA
% of molecules predicted	96.4	68.5
% of organic molecules predicted	99.6	98.5
% of organometallic molecules predicted	94.6	51.6

<sup>a</sup>The percentages refer to the set of 10,000 molecules randomly extracted from CSD09 used in our tests.

**Table 6. Comparison of COSMOS and CORINA in Terms of Speed (Mean Time Per Prediction) and Accuracy (Rmsd Value and Percentage of Molecules with Atom Clashes)<sup>a</sup>**

	COSMOS	CORINA	CSD
common set (6,795 molecules)			
time (s)	0.151	0.134	
rmsd (Å)	1.566	1.600	
clash %	14.1	18.0	0.2
% of molecules with rmsd <0.3	12.2	10.3	
% of molecules with rmsd <1.0	37.0	35.4	
organic (3,538 molecules)			
time (s)	0.099	0.182	
rmsd (Å)	1.258	1.133	
clash %	6.6	4.0	0.1
% of molecules with rmsd <0.3	17.4	18.5	
% of molecules with rmsd <1.0	46.4	54.4	
metal–organic (3,257 molecules)			
time (s)	0.208	0.082	
rmsd (Å)	1.899	2.108	
clash %	22.2	33.3	0.3
% of molecules with rmsd <0.3	6.6	1.4	
% of molecules with rmsd <1.0	26.8	14.7	

<sup>a</sup>The results are broken down by set of molecules: (1) molecules where both methods generated a 3-D structure, (2) subset of organic molecules, and (3) subset of organometallic molecules. The CSD column reports the percentage of original CSD molecules with atom clashes.

**Figure 10.** Histograms of the rmsd values obtained from the 3-D structures predicted with COSMOS (green) and with CORINA (red).

0.3 Å and (2) lower than 1.0 Å. The first measure was used in<sup>15</sup> as a valid upper limit to assess the accuracy of a prediction: if two

**Table 7. Comparison of COSMOS and CORINA for the Reduced Set of Molecules with Low Similarity with Molecules in the “Training Set” of COSMOS (CSD04)<sup>a</sup>**

	COSMOS	CORINA	CSD
common set (4,710 molecules)			
rmsd (Å)	1.729	1.728	
clash %	17.5	20.9	0.3
organic (2,163 molecules)			
rmsd (Å)	1.384	1.207	
clash %	8.4	4.6	0.1
metal–organic (2,547 molecules)			
rmsd (Å)	2.022	2.171	
clash %	25.2	34.7	0.4

<sup>a</sup>The CSD column reports the percentage of original CSD molecules with atom clashes. The results are broken down by set of molecules as in Table 6.

**Table 8. Comparison of COSMOS and CORINA in Terms of Speed (Mean Time Per Prediction) and Accuracy (rmsd Value and Percentage of Molecules with Atom Clashes) for Rigid Segments Only<sup>a</sup>**

	COSMOS	CORINA	CSD
common set (9,027 rigid segments)			
time (s)	0.073	0.097	
rmsd (Å)	0.530	0.664	
clash %	5.2	6.0	0.1
organic (3,489 rigid segments)			
time (s)	0.070	0.184	
rmsd (Å)	0.323	0.296	
clash %	1.2	1.3	0.0
metal–organic (5,538 rigid segments)			
time (s)	0.075	0.042	
rmsd (Å)	0.660	0.896	
clash %	7.6	9.0	0.1

<sup>a</sup>The results are broken down by set of fragments: (1) entire set of rigid segments, (2) subset of organic rigid segments, and (3) subset of organometallic rigid segments. The CSD column reports the percentage of original CSD rigid segments with atom clashes.

structures have rmsd < 0.3 Å, they can be considered equivalent conformations. In our evaluation both COSMOS and CORINA fail to produce a high percentage of models below this stringent threshold. The threshold of 1.0 Å provides a more forgiving measure for considering a prediction successful. Using both thresholds, CORINA is slightly better on the organic subset, and COSMOS is significantly better on the metal–organic subset. It is important to note that, in practice, rmsd is length dependent; thus, these arbitrary cutoff values are only informative when considering large sets of predictions. For instance, for a target molecule of eight atoms, an rmsd value of 1.0 Å would represent a very poor prediction, whereas a result of 1.0 Å for a molecule of 150 atoms would be quite good. To provide a sense of how the rmsd values distribute overall, Figure 10 compares the histograms of the rmsd values calculated from the predicted models of CORINA and COSMOS.

In Table 7, we also show the accuracy results for a reduced set of 4,710 molecules. This test set was obtained by removing from

**Table 9.** Comparison of COSMOS and CORINA in Terms of Correctness of Metal Coordination Geometries Using the Monocentric Fragments Extracted from CSD09Test<sup>a</sup>

	$\overline{\Delta\theta} < 10.0^\circ$			$\overline{\Delta\theta} < 15.0^\circ$		
	CSD	COSMOS	CORINA	CSD	COSMOS	CORINA
coordination number 3						
trigonal planar	111	70.3%	79.3%	133	77.4%	82.0%
t-shaped	41	26.8%	17.1%	61	41.0%	14.8%
trigonal pyramidal	73	64.4%	41.1%	99	64.6%	46.5%
unclassified	190	57.4%	22.1%	122	73.8%	51.6%
total	415	59.0%	40.2%	415	68.0%	54.7%
coordination number 4						
square planar	647	42.7%	0.0%	698	48.7%	1.0%
tetrahedral	872	86.7%	86.8%	1061	88.7%	93.7%
unclassified	483	45.5%	25.7%	243	61.3%	39.9%
total	2,002	62.5%	44.0%	2,002	71.4%	54.8%
coordination number 5						
square pyramidal	294	46.3%	3.7%	369	50.9%	6.5%
trigonal bipyramidal	214	52.6%	80.4%	363	55.1%	86.8%
unclassified	319	33.9%	25.1%	95	56.8%	38.9%
total	827	43.2%	31.8%	827	53.4%	45.5%
coordination number 6						
octahedral	1,212	82.4%	94.4%	1,406	85.6%	97.7%
trigonal prismatic	16	25.0%	0.0%	71	26.8%	2.8%
unclassified	308	44.5%	44.5%	59	55.9%	45.8%
total	1,536	74.2%	83.4%	1,536	81.7%	91.3%
coordination number 7						
capped octahedron	33	9.1%	NA	213	48.4%	NA
capped trigonal prism.	11	27.3%	NA	32	40.6%	NA
pentagonal bipyramidal	135	25.9%	NA	178	33.7%	NA
unclassified	372	42.7%	NA	128	49.2%	NA
total	551	36.3%	NA	552	43.4%	NA
coordination number 8						
cubic	7	57.1%	NA	22	45.5%	NA
dodecahedral	164	30.5%	NA	194	53.1%	NA
square antiprismatic	124	32.3%	NA	137	35.8%	NA
unclassified	388	49.5%	NA	330	67.9%	NA
total	683	41.9%	NA	683	56.5%	NA

<sup>a</sup> The CSD column indicates the number of monocentric fragments of each geometry, and the COSMOS and CORINA columns indicate the percentage of the corresponding predicted model fragments with the correct geometry. The criteria used to classify the fragment geometries are described in the Assessment Methodology section. For CN = 7 and CN = 8, no results are available because CORINA does not handle CN > 6.

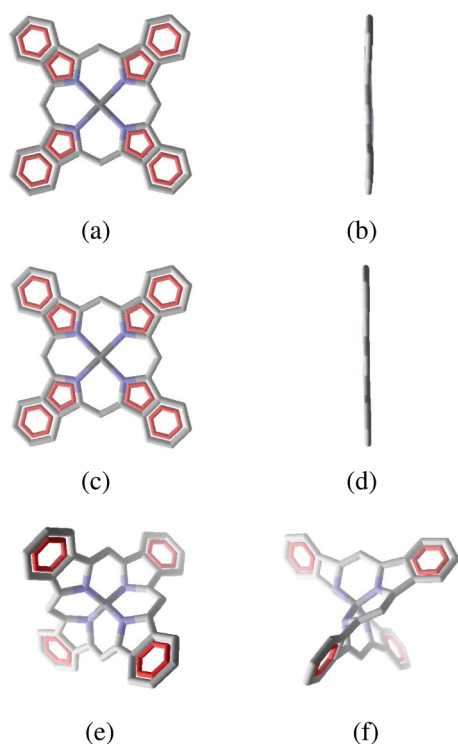
the set of Table 6 all the molecules with high (larger than 0.6) Tanimoto similarity with our “training set” (CSD04). As expected, COSMOS performs slightly worse on the redundancy reduced set; however, the results on this more difficult set still compare favorably with CORINA.

The experimentally determined structure of a small molecule in the CSD represents one low-energy conformation, but when the molecule has rotatable bonds, it may have many conformations that are approximately equal energetically. This introduces some noise in the rmsd assessment protocol. Thus, to further assess the two methods and remove a major source of ambiguity, we also compared their performance on rigid fragments in isolation. The accuracy and speed results of COSMOS and CORINA calculated on the first occurrence of each unique rigid fragment of CSD09Test are summarized in Table 8. Overall COSMOS performs better on the rigid fragment set than CORINA in terms of accuracy (0.530 Å vs

0.664 Å); however, CORINA is slightly more accurate on the organic rigid fragments, with a clash percentage (1.3%) comparable to COSMOS (1.2%). COSMOS is more accurate on the organo-metallic fragments (0.660 Å vs 0.896 Å) and is faster (0.073s vs 0.097s) overall than CORINA.

Even when rigid segments are evaluated in isolation, the rmsd is not the most appropriate measure for assessing some aspects of the models. To assess the correctness of the metal coordination geometries of the predicted models, we use the mean angle difference ( $\overline{\Delta\theta}$ ) described in detail in the Assessment Methodology section. Here, the metric is applied to the monocentric fragments of CSD09Test, and we assess how often COSMOS and CORINA models have the correct metal coordination geometry. Note that monocentric fragments which are extracted from metal-catenate structures and have incomplete metal coordination spheres, are excluded from this assessment.





**Figure 11.** Square planar versus tetrahedral modeling of a metal–organic rigid fragment. (a): Rigid fragment from the CSD with square planar metal coordination geometry. (b): CSD rigid fragment rotated 90° degrees on the vertical axis. (c): Model predicted by COSMOS with correct square planar metal coordination geometry. (d): COSMOS model rotated 90° degrees on the vertical axis. (e): Model predicted by CORINA with incorrect tetrahedral metal coordination geometry. (f): CORINA model rotated 90° degrees on the vertical axis.

We initially chose  $\overline{\Delta\theta} < 10.0^\circ$  as a threshold to classify a fragment geometry as correct on the basis of the results from comparing the ideal templates using  $\overline{\Delta\theta}$  (Table 3); however, when  $\overline{\Delta\theta} < 10.0^\circ$  is used for classification, 27.5% of the CSD fragments in the test set remain unclassified. When a more lenient threshold of  $\overline{\Delta\theta} < 15.0^\circ$  is used, only 11.2% of the CSD fragments remain unclassified. Thus, the evaluation was performed using both thresholds and the results, broken down by coordination number (CN) and specific geometry, are summarized in Table 9. Overall, the COSMOS model fragments have the correct metal coordination geometries more often than those of CORINA, using  $\overline{\Delta\theta} < 10.0^\circ$  (62.6% vs 54.0%) and  $\overline{\Delta\theta} < 15.0^\circ$  (71.3% vs 64.7%). For both thresholds, COSMOS is more accurate for CN = 3, CN = 4, and CN = 5, while CORINA is more accurate for CN = 6. When considering specific geometries, one result which stands out is that for CN = 4 CORINA predicts none of the 648 square planar fragments correctly using  $\overline{\Delta\theta} < 10.0^\circ$ , while COSMOS predicts 42.7% correctly. One example of a rigid fragment where the COSMOS model correctly utilizes square planar geometry and the CORINA model incorrectly utilizes tetrahedral geometry is shown in Figure 11. Similarly, for CN = 5, CORINA only predicts 3.7% of the square pyramidal fragments correctly using  $\overline{\Delta\theta} < 10.0^\circ$ , while COSMOS predicts 46.3% correctly. These results indicate that CORINA is biased to use certain metal coordination geometries and avoid others. While this type of bias may have a negative effect on accuracy in

some cases, CORINA's bias toward using octahedral geometry does not hurt the overall accuracy for CN = 6 because it is the correct geometry for the overwhelming majority of the structures in the test set.

Overall, COSMOS compares favorably with CORINA in terms of the correctness of the metal coordination geometries. However, these results demonstrate that both methods have room for improvement, and they give a sense of the difficulty of predicting the correct metal geometry when there are multiple plausible geometries. Because CORINA does not handle CN > 6, only the COSMOS results are presented for CN = 7 and CN = 8. For CN = 7, 36.2% of the COSMOS model fragments have the correct geometry using  $\overline{\Delta\theta} < 10.0^\circ$ , and 43.3% are correct with  $\overline{\Delta\theta} < 15.0^\circ$ . For CN = 8, 41.8% of the COSMOS model fragments are correct using  $\overline{\Delta\theta} < 15.0^\circ$ , and 56.6% are correct with  $\overline{\Delta\theta} < 15.0^\circ$ .

## CONCLUSION

We have reviewed the field and described a general framework for the rapid generation of accurate 3-D coordinates for small molecules that relies primarily on precomputed libraries of fragments, as opposed to rules or physics approximations. The power law distributions of the rigid segments and torsion angle data found in current databases of small molecules indicates that a very good coverage of chemical space can be achieved with relatively small, precomputed, libraries and that adding new fragment information will help to increase the coverage of rarely occurring fragments and torsion angles. The framework was deployed and tested here using fragment parameters obtained only from the CSD04 to predict structures of novel molecules in the CSD09. In the benchmark tests using these fragments, the performance of COSMOS and CORINA were comparable, with a clear advantage for COSMOS on metal–organics. The current version of COSMOS is available for academic research through the ChemDB chemoinformatics Web portal at <http://cdb.ics.uci.edu/> as a Web server.

While the current version of COSMOS provides an improvement over the current state-of-the-art predictions methods, it does not yet fully achieve all the specifications that we believe to be desirable for the next generation of prediction systems. We believe that further improvements of COSMOS are possible in many ways, for instance, by modifying the libraries and by developing more complete representations of both the fragment and the torsion angle distributions.

We also hope that future versions of COSMOS, or other similar systems, will be able to achieve the greater degree of data and software openness that is indispensable for real scientific progress in the field. One obstacle in this area may be the closed nature of the CSD, which unlike the PDB cannot be used without severe restrictions, even for academic research purposes. This is yet another example of the unfortunate state of affairs in chemoinformatics, where an overly zealous culture of closeness and secrecy, sometimes related to short-term profits, have greatly hampered scientific progress. A culture shift toward open access to critical data, algorithms, and software derived from them—as is pervasively the case in physics or biology—is essential to improve the quality of the science and ultimately lead to technological advances of real economic value that benefit the largest number of taxpayers.

## ■ AUTHOR INFORMATION

## Corresponding Author

\*E-mail: pfbaldi@ics.uci.edu.

## Present Addresses

<sup>||</sup> Currently with Applied Proteomics, Glendale, CA.

## Author Contributions

<sup>§</sup> These authors contributed equally.

## ■ ACKNOWLEDGMENT

Our work is partly supported by a NIH Biomedical Informatics Training grant (LM-07443-01), NSF MRI grant (EIA-0321390), and NSF grant IIS-0513376 to P.B., by a NIH/NLM Pathway to Independence Award (K99LM010821) to A.R., and by the UCI Institute for Genomics and Bioinformatics. We thank Ramzi Nasr and Jordan Hayes for helping with the calculation of the Tanimoto coefficients and with setup of the Web server. We also acknowledge the OpenBabel project and OpenEye Scientific Software for their free software academic license.

## ■ REFERENCES

- (1) Irwin, J. J.; Shoichet, B. K. ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 177–182.
- (2) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: A public database of small molecules and related cheminformatics resources. *Bioinformatics* **2005**, *21*, 4133–4139.
- (3) Chen, J.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. ChemDB update: Full text search and virtual chemical space. *Bioinformatics* **2007**, *23*, 2348–2351.
- (4) Wang, Y.; Xiao, J.; Suzek, T.; Zhang, J.; Wang, J.; Bryant, S. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.
- (5) Lauri, G.; Bartlett, P. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51–66.
- (6) Taylor, R. Life-science applications of the Cambridge Structural Database. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 879–888.
- (7) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.
- (8) Allen, F. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- (9) Allen, F.; Motherwell, W. Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 407–422.
- (10) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
- (11) Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.
- (12) Brooks, B.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (13) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **1999**, *151*, 283–312.
- (14) Case, D.; Cheatham, T., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (15) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Model.* **1994**, *34*, 1000–1008.
- (16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (17) Simons, K.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209–225.
- (18) Chivian, D.; Kim, D.; Malmstrom, L.; Bradley, P.; Robertson, T.; Murphy, P.; Strauss, C.; Bonneau, R.; Rohl, C.; Baker, D. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **2003**, *53*, 524–533.
- (19) Zhang, Y.; Kolinski, A.; Skolnick, J. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **2003**, *85*, 1145–1164.
- (20) Wishart, D.; Knox, C.; Guo, A.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
- (21) Wishart, D.; Knox, C.; Guo, A.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36* (Database Issue), D901–D906.
- (22) Tabassum, S.; Pettinari, C. Chemical and biotechnological developments in organotin cancer chemotherapy. *J. Organomet. Chem.* **2006**, *691*, 1761–1766.
- (23) Berger, I.; et al. In vitro anticancer activity and biologically relevant metabolism of organometallic ruthenium complexes with carbohydrate-based ligands. *Chem.–Eur. J.* **2008**, *14*, 9046–9057.
- (24) Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham, T.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER: A package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.
- (25) Crippen, G. M. Rapid calculation of coordinates from distance matrices. *J. Comput. Phys.* **1978**, *26*, 449–452.
- (26) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformations*; Chemometrics Series; Wiley: New York, 1988.
- (27) Pearlman, R. Rapid generation of high quality approximate 3-D molecular structures. *Chem. Des. Auto. News* **1987**, *2*, 1–6.
- (28) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3-D structure of a molecule in its IR spectrum. *Fresenius J. Anal. Chem.* **1997**, *359*, 50–55.
- (29) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Model.* **2003**, *43*, 1186–1191.
- (30) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: A FRee Online druG 3-D conformation generator. *Nucleic Acids Res.* **2007**, *35*, W568–572.
- (31) Pearlman, R. S. In *3-D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 41–79.
- (32) Sayers, E.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2010**, *38*, D5–D16.
- (33) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. L. The Blue Obelisk: Interoperability in chemical informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
- (34) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 15869–15872.
- (35) Gasteiger, J.; Engel, T., Eds. *Chemoinformatics: A Textbook*; Wiley: New York, 2003.
- (36) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, The Netherlands, 2005.
- (37) Weininger, D.; Weininger, A.; Weininger, J. SMILES: A chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

- (38) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (39) Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237–243.
- (40) Wheeler, D.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2007**, *35*, D5–D12.
- (41) Lang, P.; Brozell, S.; Mukherjee, S.; Pettersen, E.; Meng, E.; Thomas, V.; Rizzo, R.; Case, D.; James, T.; Kuntz, I. DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA* **2009**, *15*, 1219–1230.
- (42) Baldi, P.; Benz, R. W.; Hirschberg, D.; Swamidass, S. Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.
- (43) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754, PMID: 20426451
- (44) Gasteiger, J. Chemoinformatics: A new field with a long tradition. *Anal. Bioanal. Chem.* **2006**, *384*, 57–64.
- (45) Crabtree, R. H. *The Organometallic Chemistry of the Transition Metals*; Wiley: New York, 1988; pp 32–33.
- (46) Benz, R. W.; Swamidass, S. J.; Baldi, P. Discovery of power-laws in chemical space. *J. Chem. Inf. Model.* **2008**, *48*, 1138–1151.
- (47) Feuston, B.; Miller, M.; Culberson, J.; Nachbar, R.; Kearsley, S. Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 754–763.
- (48) Bondi, A. Van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451.
- (49) Sadowski, J. In *Molecular Drug Properties: Measurement and Prediction*; Mannhold, R., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; pp 157–181.
- (50) Wipke, W.; Hahn, M. Analogy in computer-assisted design. *Prog. Clin. Biol. Res.* **1989**, *291*, 141–145.
- (51) Cordero, B.; Gómez, V.; Platero-Prats, A.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, *21*, 2832–2838.
- (52) Halgren, T. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.