JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

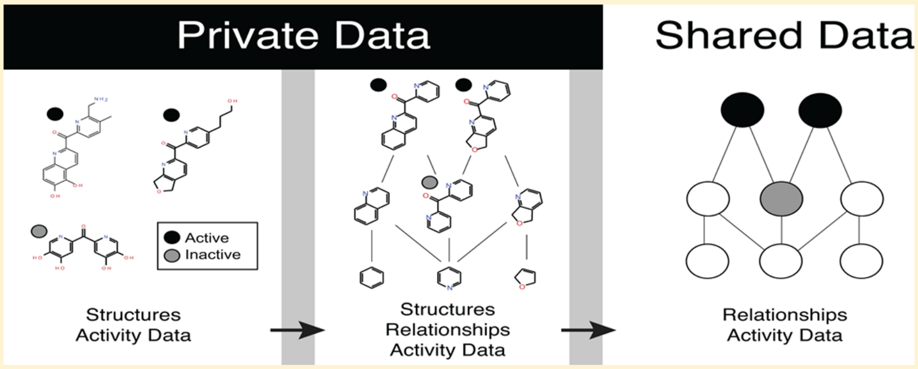
pubs.acs.org/jcim

Sharing Chemical Relationships Does Not Reveal Structures

Matthew Matlock and S. Joshua Swamidass*

Washington University School of Medicine, Department of Pathology and Immunology, St. Louis, Missouri 63110, United States

S Supporting Information



ABSTRACT: In this study, we propose a new, secure method of sharing useful chemical information from small-molecule libraries, without revealing the structures of the libraries' molecules. Our method shares the relationship between molecules rather than structural descriptors. This is an important advance because, over the past few years, several groups have developed and published new methods of analyzing small-molecule screening data. These methods include advanced hit-picking protocols, promiscuous active filters, economic optimization algorithms, and screening visualizations, which can identify patterns in the data that might otherwise be overlooked. Application of these methods to private data requires finding strategies for sharing useful chemical data without revealing chemical structures. This problem has been examined in the context of ADME prediction models, with results from information theory suggesting it is impossible to share useful chemical information without revealing structures. In contrast, we present a new strategy for encoding the relationships between molecules instead of their structures, based on anonymized scaffold networks and trees, that safely shares enough chemical information to be useful in analyzing chemical data, while also sufficiently blinding structures from discovery. We present the details of this encoding, an analysis of the usefulness of the information it conveys, and the security of the structures it encodes. This approach makes it possible to share data across institutions, and may securely enable collaborative analysis that can yield insight into both specific projects and screening technology as a whole.

INTRODUCTION

We propose sharing the *relationships* between molecules along with assay and property data (Figure 1 and 2). The idea is to share assay data associated with each molecule, to share information about which molecules are related one to another in the data set, and to withhold any additional information about their structures. This proposal contrasts with prior approaches to securely sharing chemical information, which focused on sharing descriptors of chemical structure. Structural descriptors can often be reverse engineered into structures, making them insecure to share. In contrast with structural descriptors, the relationships between structures are much more difficult to reverse engineer.

It might seem as though the relationships between chemicals are not useful. However, most algorithms and methods in chemical informatics do not directly use structural descriptors. Instead, in their first step, they convert structural descriptors into some representation of relationships between molecules, and it is through these relationships that structural inferences are made. For example, fingerprint similarity (based on combinatorial generated structural descriptors) is often used to predict the properties of molecules. The first step in these predictions is using a query molecule's fingerprint (a structural description) to compute similarities to molecules with known properties (a representation of relationships between molecules). The structural descriptions are not important if the relationships are known. Most methods in chemical informatics work in this model. Our approach enables secure analysis of chemical structures any method that works in this model.

Secure analysis of chemicals is important for several reasons. First, over the last several years, many new ways of analyzing screening data have been developed. There is nonobvious, valuable information in screening data, and new methods of extracting this information are actively being developed. To benefit from these methods, screening organizations need to either develop in house implementations of these methods or collaborate directly with the researchers that are developing

Received: July 8, 2013
Published: December 1, 2013

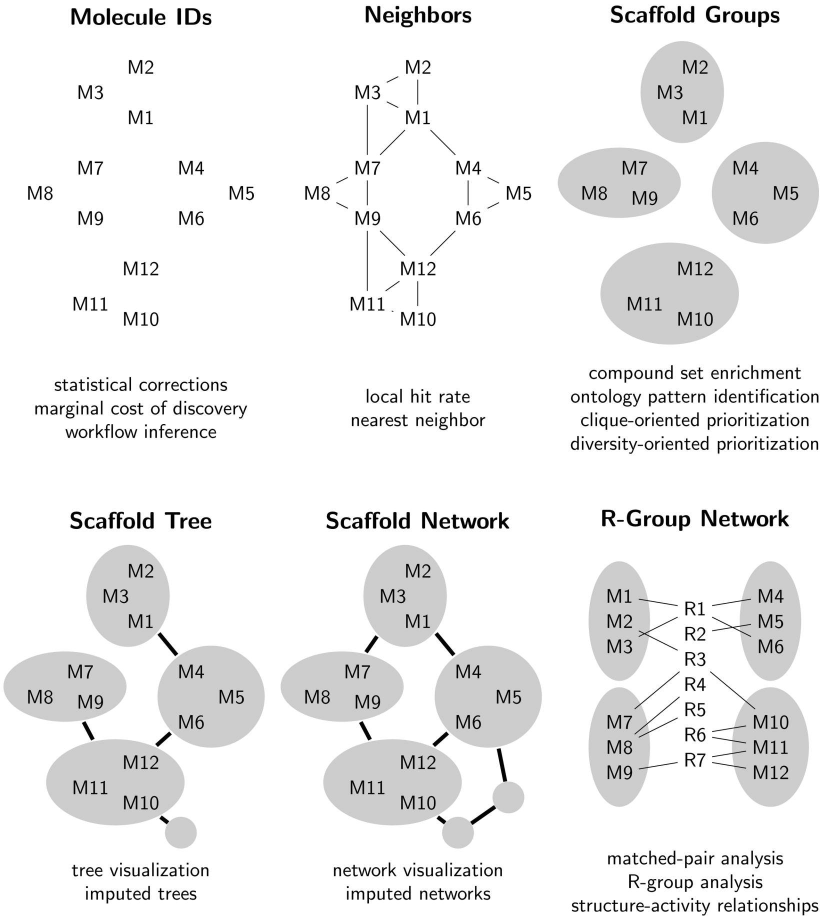


Figure 1. We propose six ways of sharing the relationships between molecules, with increasing amounts of useful chemical information. Some of the analyses that are possible with each type of data are listed below each option. First (top left), we propose sharing molecule IDs with assay data. In this case, no information about chemical structures are shared, but this is still enough information to perform analyses that do not depend on structure. Second (top middle), we could share which molecules are structure neighbors. Third (top right), we could group molecules together if they share a common scaffold. Fourth (bottom left), we could share how these scaffolds connect to one another in a tree. Fifth (bottom middle), we could share how these scaffolds connect together in a network. Both the tree and network include empty scaffold groups, which are shown in the figure. Sixth (bottom right), we could group molecules into scaffold groups, and annotate each molecule with the R-groups they contain.

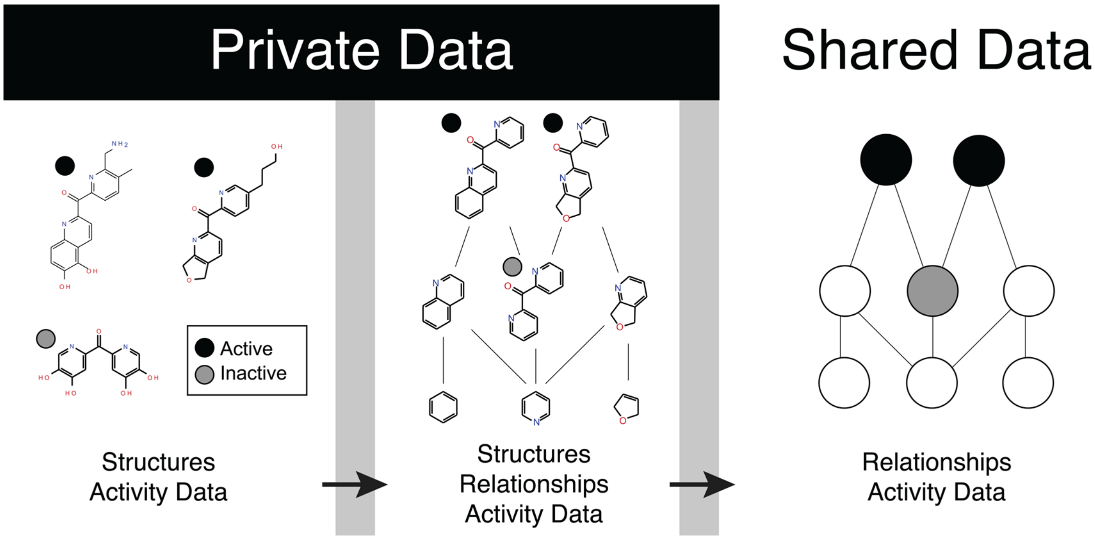


Figure 2. Sharing relevant information by sharing relationships between molecules instead of structural descriptors. Internally, one party holds private data that consists of molecule structures associated with assay data. The relationships between molecules in this data can be shared without revealing molecule structures. These relationships are informative enough to use several important chemical informatics algorithms on screening data. In contrast, structures can be reverse engineered from most structural descriptors.

these methods. Finding ways to share relevant chemical information about screening data that leaves structures blinded could open the door for this type of collaborative work. These methods include better strategies for identifying active molecules from primary screens, which leverage information from fingerprints,[1] scaffold groupings,[2,3] economic modeling,[4−6] and improved processing of raw data.[7−9] They also include automatic methods of organize screening data into workflows,[10] a series of powerful strategies for visualizing how biological activity maps to chemical space,[11−14] and understanding the contribution of specific substitutions to the properties of the molecule.[15]

Second, secure methods of sharing chemical information could make outsourcing and collaboration over screening data possible. Outsourcing is increasingly important in drug discovery because it reduces the cost of many Research and Development efforts and enables centralization of expertise.[16−18] In the same way this change has improved experimental efforts, outsourcing could also reduce the cost and, perhaps, increase the value derived from the analysis of assay data.

Third, if more data is made available, enabled by secure methods like ours, it is possible that unexpected connections and patterns in data could be identified. Of particular interest are unexpected signals in screening data which indicate either specific molecules or mechanisms by which to treat human disease[19] and the identification of promiscuous molecules that can waste substantial resources in drug development projects.[20] As more and more data is collected, it can often be used for purposes other than those for which it was originally collected. Sharing large collections of proprietary assay data, with structures blinded, would enable researchers not involved in the original data collection to identify important molecules and mechanisms to treat human disease.

Information Barrier to Sharing Chemical Descriptors. The 2005 American Chemical Society meeting included a session on securely sharing chemical information to support collaborative development of ADME predictors.[21] The papers based on the presentations in this session were published in a special issue of the *Journal of Computer-Aided Molecular Design*.[22−31] Little progress toward securely sharing molecular data has been made since they were published eight years ago, so these papers remain an excellent review of the currently prevailing views on this topic.

The most important result from this body of work, independently derived by two groups, is that the information content in molecular structures is as low as 1 bit per atom.[26,29] This is a result from information theory, which was arrived at by empirically measuring the compression ratios of structures in SMILES format by commonly used compression software.[32,33] While this result does not generate an algorithm for reverse engineering molecules from their descriptors, it strongly suggests that such algorithms exist for any descriptor or combination of descriptors that contains more than about 1 bit of information per atom. In support of this result and its interpretation, at least three methods are reported which demonstrate that structures can be recovered from commonly used descriptors.[22,26,34] Perhaps surprisingly, one of these methods, on the basis of de novo design software, was able recover structures from a single, computed logP value.[22]

This result did not stop other groups from proposing methods for sharing chemical data securely. One group suggested topological indexes are very hard to reverse engineer if they are degenerate, associating the same number to large numbers of molecules.[23] Topological indexes are not that useful for the prediction of biological activity, so this result does not seem to be consequential. One group suggested a software design solution and data file format for sharing descriptors,[25] but this neglects the core of the problem. Which descriptors can be shared that do not give away too much information?

One approach is to generate less informative descriptors. One group suggested increasing the degeneracy of 2D fingerprints to enable useful but noisy measures of similarity.[24] Another group suggested sharing the similarity of hidden molecules to a set of reference molecules.[27] Another group suggested that sharing degenerate descriptors based on substructures could be safe, but also noted that the risk can increase as several descriptors are shared.[31,35] Unfortunately, none of these approaches preclude the use of de novo design programs to reverse engineer structures because as little as 1-bit of information per atom creates a real risk to exposing structures.[22,26,34] The fundamental problem with these approaches is that the amount of information shared must be so low that it becomes questionable whether or not it would be useful for discovering novel drug compounds.

Another approach is to share descriptors of something other than the structures being hidden. One group suggested sharing descriptors of groups of structures instead of individual structures.[28] In a related approach, another group suggested using surrogate structures, by picking structural neighbors of the structures that are being hidden, to generate shareable descriptors.[29] Both approaches are promising because they break the direct link between the structures to be hidden and the shared information, but by their very nature they still enable attackers to determine the class of structures to which the hidden molecules belong.

Another approach is to enable secure, collaborative computation across multiple sites.[30] This approach uses ideas from cryptography, which rely on the difficulty of factoring large integers, to enable secure matrix multiplications by multiple parties that each privately hide specific portions of the matrix. For our purposes, the matrix holds the descriptors of hidden molecules, and the matrix multiplications enable researchers to perform a collaborative regression across all the molecules to any property of interest. This approach has some promise, but has some important drawbacks. First, making matrix multiplication available enables a malicious party to recover all the private matrix elements by requesting a secure multiplication between the private data and an identity matrix. Second, it requires a substantial amount of software overhead and message passing between different parties. Both problems make this approach impractical for most applications in chemical informatics.

The takeaway from this body of work was that ingenuity cannot overcome the cold, mathematical fact that chemical structures have very low information content, so it may not be possible to share relevant information about individual molecules or molecule libraries without revealing the details of their structure.[26]

Since this session at ACS, the desire to share chemical data has not diminished, and there have been some notable efforts toward sharing chemical information. For example, the Lilly Open Innovation Drug Discovery program seeks to find people willing to share physical chemical samples to test in phenotypic assays, after intellectual property agreements are signed. Collaborative Drug Discovery (CDD) developed a secure

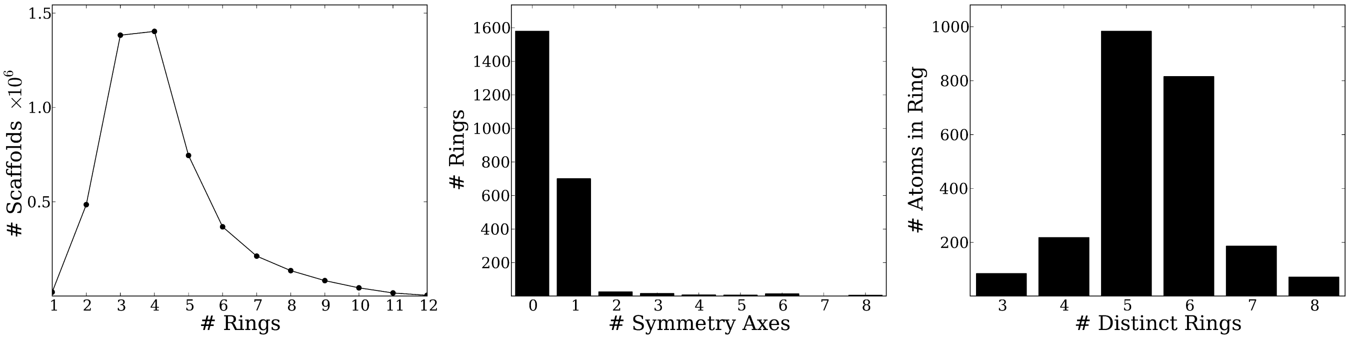


Figure 3. PubChem data. (Left) the distribution of scaffolds in the PubChem data set by number of rings. (Center) distribution of rings from PubChem over their number of symmetry axes. Most rings are asymmetric. (Right) the distribution of rings from PubChem over the number of atoms in the ring.

web portal within which groups can securely share chemical structures with each other.[36] With Pfizer and affiliates of CDD, one group suggested using the standards implemented in open source software as a basis for sharing QSAR models.[37] In this case, descriptors of molecule are not shared, but the rules used to predict the activity of molecules based on open-source descriptors are shared. The NIH launched a drug repurposing program, where clinical and assay data about failed drugs is shared with academic researchers, once again after agreements are signed and without revealing private structural information.[38]

These efforts are evidence that there remains a persistent desire to share chemical information. At the same time, no one is openly sharing structural descriptors of the molecules in private collections; structural data is only shared after intellectual property agreements are in place. This is for good reason. As we have seen, doing so would expose these structures and is, therefore, not secure.

We aim to develop novel strategies for openly sharing information about private collections of molecules in a way that does not reveal their structures. These strategies would supplement and enhance current efforts to collaborate and to understand drug discovery as a whole.

Overcoming the Information Barrier by Sharing Relationships. The theoretical barrier to sharing this type of chemical information is the low information content in individual molecule structures. This is a fundamental barrier, rooted in information theory, that must be circumvented by any approach that hopes to securely share relevant information. At first, this barrier seems insurmountable. However, it is possible to circumvent it by sharing the *relationships* between molecules in a data set, instead of descriptors of molecule structures.

This is the key innovation of our approach, and it has three key consequences. First, this approach shares the information required for most algorithms to analyze and visualize screening data, so it is useful information. Most of these methods use structural descriptors to compute the relationships between molecules in the data set, after which the descriptors are set aside. By moving this computation to the private space, we can hide descriptors to prevent them from being used to reverse engineer molecules.

Second, from the perspective of information theory, less information exists in the relationships between molecules than exists in their structures. This provides strong theoretical backing for the security of our approach. In the same way that the low information content in molecular structures undercuts their use to securely share chemical information, it also helps support the security of sharing the relationships between molecules.

Third, this approach makes the shared data set secure by rendering it unlinkable to other databases, either public or private. Only extremely degenerate information about structures is shared, if any information is shared at all, so it becomes impossible to measure how similar a molecule with known structure is to molecules in the private database. Certainly, one could compute a scaffold network from a molecule with known structure, but it is impossible to know how the nodes in this known network map to nodes in the unknown network.

This final consequence is a critically important feature; rendering the shared data unlinkable makes all reported approaches to reverse engineering structures unworkable. The best known strategy for reverse engineering descriptors works by (1) generating candidate molecules with known structures, (2) checking to see if their descriptors are similar to the hidden structures' descriptors, and (3) modifying the candidates to improve the match between their descriptors and those of the hidden structures.[22,26,34] Steps 2 and 3 become impossible because the descriptors of the hidden structures are not shared.

A concrete example demonstrates how this approach works in practice. A scientist with high-throughput screening data can share the data with a collaborator, sharing the assay readouts and metadata paired with IDs for each molecule. Relationship information is also shared along with this data, and this data is useful enough to enable useful analysis. If molecules are grouped into structurally related families and these groups are shared, the collaborator can use, for example, compound set enrichment analysis to identify latent hit series:[3] active scaffolds missed in the initial study. There is not enough information for the collaborator to figure out the structures of the molecules in the assay. At the same time, the collaborator can communicate back which molecules should be pulled for follow up studies using the molecule IDs. This example uses compound set enrichment, but most cheminformatic algorithms can be applied in this way, though some may require different types of relationship information to be shared.

Now, we turn to a detailed description and analysis of our approach.

■ DATA

We used publicly available chemical assay data from the PubChem database[39] to explore our approach. This database includes biochemical assays of millions of compounds. We randomly selected a large subset of the compounds available in

PubChem (see Supporting Information), resulting in a representative set of approximately 10 million compounds. After generating scaffold networks for these compounds, approximately 5 million unique scaffolds were identified. The scaffolds generated from this data set have diverse structures, containing approximately 2300 distinct rings from 3 to 9 atoms in size. In this data set, most scaffolds have from three to five rings, most rings are asymmetrical and most rings have 5 or 6 atoms (Figure 3).

■ METHODS

In this study, we propose several methods and types of information as candidates for sharing data about screening assays. These methods are evaluated in a theoretical way, using information theory, and by an empirical analysis of the types of information revealed, which structures are the most vulnerable to potential attack, and how some of these risks may be mitigated.

Scaffolds, Trees, Networks, and R-groups. First, we propose sharing molecule IDs with assay data, entirely blinding all structural information. Second, we propose sharing how molecules in a screen are connected to one another in a screening network. Third, we propose sharing how molecules are grouped together into scaffold groups. Fourth, we propose sharing how these groups are connected into trees. Fifth, we propose sharing how these groups are connected into networks. Sixth, we propose sharing how molecules are connected together into R-group networks.

The first method we propose, based on molecule IDs, shares no structural information and is include primarily to serve as a baseline against which to compare the other methods. Nonetheless, sharing molecule IDs with assay data is more useful than might be expected. Molecule IDs provide enough information to find patterns in screening data[4−6] and to enable intelligible communication about private chemical entities. For example, the NIH repurposing program depends on sharing molecule IDs with assay and clinical data.[38]

For the other other methods, we used an implementation of scaffold groups, trees and networks, which was developed in-house. Our implementation borrows code from the Scaffold-Hunter software, but can run across the 10 million compound data set obtained from PubChem. This software works by computing the Bemis−Murcko scaffolds of all molecules in the database[40] (Figure 4). The scaffolds are the core ring system of a molecule, with all side chains (the R-groups) removed. The scaffolds are then iteratively decomposed, removing each ring one at a time to systematically generate smaller scaffolds. Scaffold trees decompose each scaffold in a single, deterministic way, based on a set of predefined rules, while scaffold networks decompose scaffolds in all possible ways.[3,12] The end result of this analysis is a set of scaffolds, R-groups, and connections between scaffolds. We represent the scaffolds and R-groups as canonical SMILES strings. To share these relationships, the SMILES strings would be removed, and an ID for each distinct scaffold and R-group would be generated. These anonymous IDs, with connections between molecules, scaffolds, and R-groups would be shared.

Scaffold Network Comparison. A scaffold network computed from a single molecule is a directed, multipartite graph. The connectivity of this graph, its topology, is a signature for the molecule. As we will see, it is a noninformative signature because very different molecules can generate identical networks. To show this, we used the open source graph software Nauty[41] to produce canonical representations of scaffold networks of every molecule in our PubChem data set.

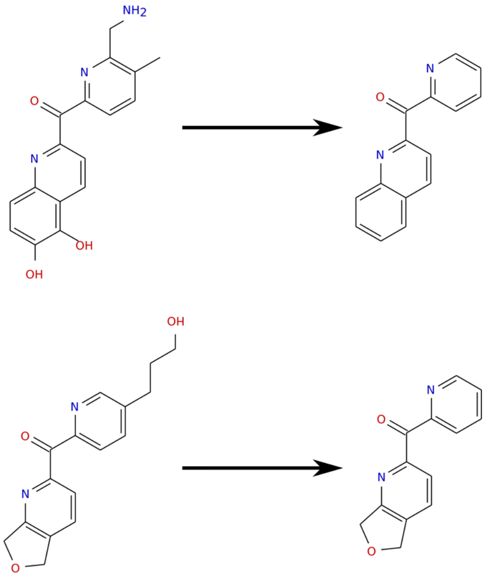


Figure 4. Scaffold for a small-molecule is obtained by removing all side chains (the R-groups), leaving the rings and linkers connecting them.

Molecule Similarity. To evaluate the diversity of compounds in a given network equivalence class, we compared each compound in an equivalence class to other molecules in that class via Tanimoto Coefficient using the default FP2 fingerprint provided by the OpenBabel Python API.[42,43]

Information Content Estimates. Using these methods, we can share the relationships between molecules in any molecular database. We can empirically approximate the amount of information, in bits, that each of these methods reveals about the data. First, we create a file containing all the data that would be shared by a method under consideration. Second, we find the most compact way possible of representing this data in a string. Third, after converting the file into the compact string representation, we compress the data using a string compression utility. The size of the compressed file puts an upper bound on the amount of information shared by the method.

■ RESULTS

In the following sections, we investigate the security of sharing information about the relationships between molecules. First, we use compression software to empirically estimate the information content of each descriptor. Second, we consider the simplest sharing methods: molecule IDs, the neighbor network, and scaffold groups. Third, we study scaffold trees in more detail. Finally, we conduct an extended analysis of the risks of sharing scaffold networks.

Information Content. The compressed size of a data set is an estimate of the amount of information in a specific data set. If the compression is reversible, the compressed size of a data set puts an upper bound on its information content. This is just an estimate, it is not exact because the compression size is increased by (1) noninformative noise in a data set and (2) any inefficiency in the compression protocol. The first confounder, as will see, will influence the estimates of information content of nearest neighbors. The second confounder is evident because some text encodings with the same information will compress to different sizes. For example, the same molecules will not compress to the same size if encoded in SDF format instead of SMILES strings. It is impossible to know for sure if the optimal encoding is being used.

Nonetheless, the compression size estimate of information content is still useful because the true information content always less than compression size. If there is a wide margin between information in the hidden structures and shared data, there is good reason to believe that algorithms to reverse engineer the data to expose the structures are not possible.

We find that all the methods of sharing relationships between molecules contain substantially less information than smiles strings (Table 1). These results are obtained by encoding the data in the most compact text representation we can invent, and then compressing this text with the bzip2 compression algorithm.[32,33] The resulting file size, in bits, is an estimate of the information content. These results are the most important result of this study, which directly overcomes the information barrier to securely sharing chemical information.

Table 1. Compression Efficiency (Measured in Bits Per Molecule, Scaffold, or Atom) Puts an Upper-Bound on the Amount of Information Shared by Each Method[a]

	mol. atom	scaff. atom	molecule	scaffold
molecule SMILES	1.49	3.57	38.58	77.64
molecule ID	0	0	0	0
R-group network	0.24	0.59	6.34	12.76
1 neighbor	0.095	0.23	2.45	4.93
3 neighbors	0.69	1.66	17.95	36.13
5 neighbors	1.29	3.09	33.39	67.20
7 neighbors	1.86	4.46	48.20	97.00
9 neighbors	2.39	5.72	61.86	124.49
scaffold SMILES	0.56	1.33	14.43	29.04
scaffold group	0.00048	0.0012	0.013	0.025
scaffold tree	0.081	0.18	1.96	3.94
scaffold network	0.34	0.82	8.91	17.93
pruned scaffold network	0.058	0.14	1.50	3.00

[a] It is not possible to reconstruct structures from sharing less information than SMILES stings. The top group of methods share per-molecule information and should be compared with the information content of molecules' SMILES strings. The bottom group of methods share per-scaffold information and should be compared with the information content of the scaffolds' SMILES strings. Most methods we propose contain less information than SMILES strings by a wide margin, suggesting that, on average, they are secure. The one exception to this is sharing more than a few structural neighbors.

For these experiments, we use the data in Assay 504333 from PubChem, data much like that from most high-throughput screens. This screen contains 350 000 molecules with an average of 25.9 atoms per molecule and about 173 000 scaffolds with an average of 21.8 atoms each. The SMILES strings from all of the molecules and just the scaffolds of these molecules were collected into two files. These files were sorted and compressed to yield the baseline information content, against which we compare each method. These methods were encoded as follows: (1) The molecule IDs contain almost zero information because all that is conveyed is the total number of molecules, a number which can be encoded in just a few characters of text. (2) Scaffold groups reveal only the number of molecules in a large number of groups. These are encoded by putting the scaffold groups' sizes in sorted order, and storing, in text, numerical difference between the sizes of consecutive groups. This is called a "run-length" encoding, and is commonly used to improve compression. (3) Scaffold trees are encoded in a notation that includes the number of molecules in each scaffold group and the structure of the tree. Here, a single character text code is used to move up or down a level in the tree. Within each level, interspersed with these codes, each scaffold is recorded as the number of molecules it contains. (4) Scaffold networks are encoded by the number of molecules in each scaffold group and a sparse adjacency matrix, which records how scaffolds are connected. In this matrix, scaffolds are reordered so scaffolds connected to one another are closer in the matrix (using the Reverse Cuthill−McKee algorithm[44]). The matrix is encoded in a skyline format to improve compression.[45,46] (5) The nearest neighbors were encoded by compressing an adjacency matrix of the whole data set. The molecules in this data set are reordered using the Reverse Cuthill−McKee algorithm to put molecules in the matrix with the same neighbors near one another, and the matrix is encoded in skyline format. (6) The R-group network is encoded by storing an integer identifier and the number of observed occurences for the R-groups associated with each molecule. Each molecule is stored on a separate line and the integer IDs are stored in sorted order and run length encoded. One more sort across the entire data set brings molecules with the same R-groups into close proximity, thereby improving the compression.

With only one exception, nearest neighbors with more than a few neighbors, all these representations compress to smaller sizes than SMILES strings by a wide margin. This is a critical result, as it demonstrates that there is a strong theoretical basis behind using these representations to share molecule data.

In most cases, there is a wide margin between the information in SMILES strings and the information we propose sharing. This provides strong support for the notion that the proposed methods of sharing chemical data are secure. This analysis only applies to the average case. It is still theoretically possible that a method concentrates its information in a narrow region of the chemical space, thereby compromising the security of a few specific molecules. Moreover, some general information about structures is conveyed by these representations. These worst-case scenarios are identified and the nature of the chemical information exposed is studied for each representation in the following sections.

Molecule IDs. It seems self-evident that sharing molecule identifiers alone is entirely secure. Of course, the assumption is that the identifiers are private, and there is no public method of relating them back to their structures. In practice, this is easily achieved by either (1) sampling a random integer to use as the ID, (2) sequentially number molecules, or (3) using an internal identifier for which their is no public method of relating them back to structures. Sharing IDs like these is safe because there is essentially zero information content, other than the size of the screening library.

Nearest Neighbors. Nearest neighbors do not compress as efficiently as we expected. Although the network constructed by sharing one neighbor is just 2.45 bits per molecule (compared with 38.58 bits per molecule in SMILES), and 3 neighbors have an information content at 17.95 bits per molecules (less that half that of SMILES), increasing the number of neighbors rapidly increases the information. There are a few explanations for this. First, neighbors could be more informative than they appear and an algorithm for reverse engineering them may exist. Second, it is possible that our encoding is far from optimal

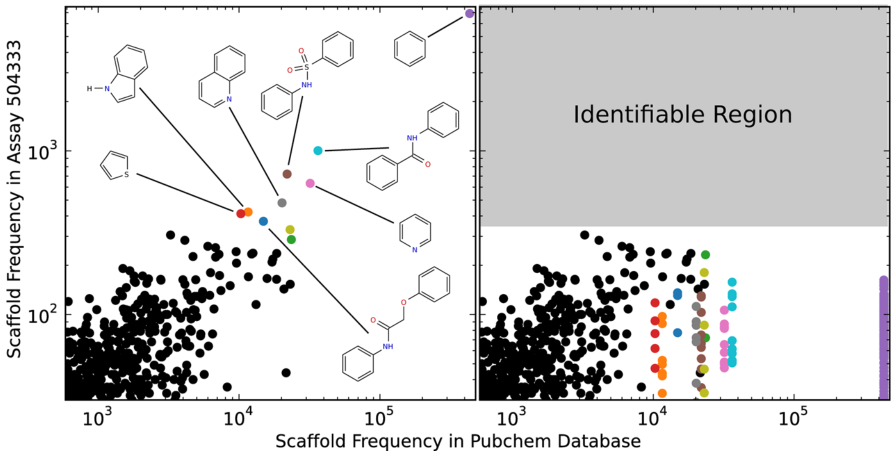


Figure 5. Most frequent scaffolds in different data sets are often the same. Plotting the top 1000 most frequent scaffolds from a specific assay (AID 504333) against the top 1000 scaffolds from our PubChem data set shows that the most frequent scaffold is usually benzene (left). This fact enables benzene to be identified from scaffold groups; it corresponds with the largest scaffold group. A small number of additional scaffolds (also shown) are vulnerable to this attack. Splitting these large scaffold groups into smaller groups breaks this relationship and counters this attack (right). The split groups are indistinguishable from the other scaffolds in the screen.

and a better ordering of the matrix could dramatically improve the compression. Third, there could be a substantial amount of uninformative noise in the neighbor matrix, which does not compress efficiently.[47]

We suspect the correct explanation is a combination of the last two theories: that there are substantial improvements possible to the encoding and that there is substantial uninformative noise in this representation. This notion is supported by the fact that the more distant neighbors increase the information content by a significantly greater margin than the closer neighbors. Clearly, the closest neighbors are most informative for the purpose of chemical analysis. At the same time, the distant neighbors have the most noise. The difficulty in compressing distant neighbors is probably more a consequence of the inherent noise introduced with these distant relationships and our difficulty encoding them, rather than the quantity of information about molecule structures that they contain. Restricting the nearest neighbors graph to only include edges between highly similar molecules (with Tanimoto > 0.765) yields additional support for this hypothesis. In this case, there are still an average of 5 neighbors per node, but the information content is only 10.6 bits per molecules, which is significant lower than the 5-nearest-neighbors (33.4 bits per molecule).

While it is probably safe to share neighborhood networks, this analysis cannot rule out the possibility that reverse engineering algorithms exist which could recover molecule structures. At the same time, we cannot imagine how an algorithm like this might work. In the future better methods of compressing the neighbor matrix might be developed, which could enable us to prove that a solution does not exist. Nonetheless, our inability to conceive of such an algorithm is not a proof of nonexistence. At the current time, in light of these uncertainties, we would recommend only sharing a few neighbors for each molecule in the database. To be sure of security risks, we also recommend comparing the number of bits being shared per molecule to the number of bits in their SMILES strings to ensure a wide security margin.

Scaffold Groups. Scaffold groups share a little bit more information than molecule IDs (only 0.01 bits per molecule), and they are nearly as secure. The only information about scaffolds being shared is the number of molecules in each group. The frequency of scaffolds is not correlated with any chemical property. For example, there is only a very weak negative correlation between the number of molecules in scaffold and its molecular weight.

However, we did find one vulnerability. Often the most common scaffolds in two different data sets are the same (Figure 5A). As one would expect, the most frequent scaffold in most screening libraries is benzene, and benzene is identifiable in the scaffold groups generated from these data sets because of this. This specific vulnerability can be mitigated by breaking up the most frequent scaffold groups into a large number of smaller groups. These splits could be random or based on keeping molecules with similar R-groups together. By splitting the frequent groups in this way, benzene becomes unidentifiable (Figure 5B). At the same time, the most frequent scaffolds are also the least interesting from a drug development perspective, so it is possible that revealing their structures is not threatening.

It is worth understanding how this vulnerability arises even when the information content of scaffold groups is so low. At face value, this vulnerability may seem to contradict our results from information theory. In reality, most of the information is concentrated on a small region of the chemical space, on the scaffolds with high frequency, which is why they can be identified.

R-Group Network. The R-group network contains less than a quarter of the information in SMILES strings (6.34 compared with 38.58 bits per molecule). This is not surprising: R-group networks carry very little information about the most important and variable portion of a molecule, its scaffolds. As might be expected, R-groups have the same vulnerability as scaffold groups. The most common R-groups are consistent across several data sets. Some of the most frequent groups are methyl groups (−C), methoxy groups(−CO), halides (−Cl, −Br, and −F), and single heteroatoms (=S, =N, −N, and −O). This vulnerability is more problematic for R-groups because there much fewer of them than scaffolds. So there is more risk that they can be exposed.

In a strategy similar to that used with scaffold groups, this vulnerability can be countered by artificially splitting nodes in the R-group network. Recall that in R-group networks, each R-group is represented as a node in a graph, and molecules that have this group as a side-chain are connected with an edge to the node. In the splitting operation, the node corresponding to a frequent R-group is split in to several nodes, and the molecules that connected to the original node are divided up across all the new nodes. The usefulness of the network can be preserved by ensuring molecules that are close in chemical space (eġthat are derivatives of the same or closely related scaffolds) are grouped together to connect to the same R-group node when the original R-group node is split. Splitting nodes in the network breaks the relationship between R-group frequency and molecular structure, and makes the network entirely secure.

Without this counter measure, commonly encountered R-groups can be exposed. At the same time, exotic, rarely encountered side chains remain unidentifiable. It also is worth considering how important it is to protect the identity of common R-groups in the network. Knowing which common side chains a molecule has reveals very little about that molecule's structure. It is not likely, for example, to be grounds for challenging or granting a patent or enough information to create a focused combinatorial library. In some scenarios, one might elect to reveal the exact structure of R-groups because they do not appear to be dangerous to share.

Scaffold Trees. Scaffold trees contain only a small amount of information: 1.96 bits per molecule, compared with 38.58 bits in SMILES strings. This result demonstrates a wide information gap between trees and molecule structures. This makes sense, because it does not seem possible to reverse engineer molecules from their scaffold trees, since these trees only expose a limited subset of scaffold relationships.

Scaffold trees convey the same information as scaffold groups, and have the same vulnerabilities. In the same way, these vulnerabilities can be countered by artificially splitting the nodes associated with the most frequent scaffolds. This ensures that, for example, the benzene scaffold cannot be identified.

Scaffold trees also convey some very general information about scaffolds. First, the level of the scaffold in the tree corresponds exactly to the number of rings in the molecule. Second, because child scaffolds always contain one more ring than their parents, the molecular weight of scaffolds loosely correlates with their level in the tree (Figure 6).

There does not appear to be much more information that can be reverse engineered from scaffold trees. Recall that scaffold trees represent scaffolds as nodes in a network. Each scaffold is decomposed by removing rings, one at a time. Every node has only a single parent node. Every scaffold is the parent of a combinatorially large number of scaffolds with one more ring. Approximately, each scaffold could have as many as

$$LRA/N \approx 1700 \cdot 2000 \cdot 25 \cdot N/N = 8.5 \text{ million} \quad (1)$$

children, where L is the number of possible linkers (about 1700 with up to 5 atoms including carbon, nitrogen, oxygen, and carbonyl groups), R is the number of rings (about 2000 of 5 to 7 atoms in PubChem), A is the number of ways the linker can attach a ring to a scaffold (about $5 \cdot N \cdot 5$), and N is the number of rings in the parent scaffold. The division by N is introduced because the scaffold tree algorithm strips the least informative ring first in each decomposition; more rings in the scaffold make less rings acceptable to be added. This is only a rough approximation, and could surely be improved by more carefully considering symmetries and how the number of attachment points grows as scaffolds increase in size. The key point that will not change with an improved estimate is that there are more than a million possible children for every scaffold in a scaffold tree. This one-to-many relationship between parent and child scaffolds is what produces the tree structure of the network, and is also what makes it difficult to reverse engineer the network into molecule structures. Even if one of the scaffolds is exposed (for example, by identifying the most frequent scaffold as benzene), there is no way to know which of the millions of possible children correspond to child nodes of benzene in a specific tree.

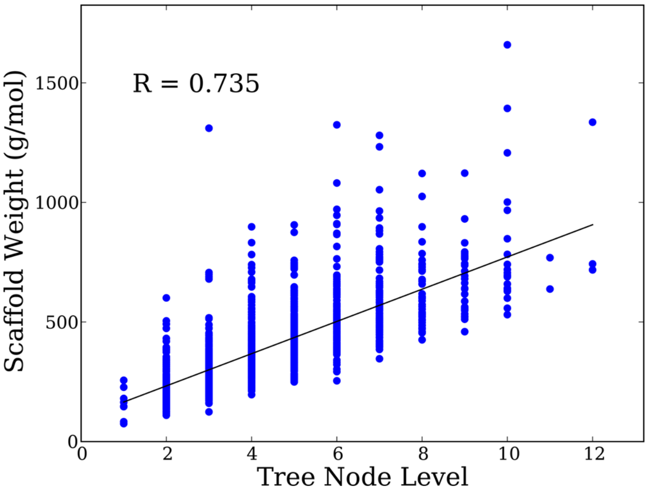


Figure 6. Molecular weight in scaffold networks and trees. The level of a scaffold in the tree or network conveys a rough sense of its molecular weight. The higher the level, the higher the weight, because higher level scaffolds have more rings.

Scaffold Networks. Scaffold networks are the most informative of the methods we propose. They contain 8.91 bits per molecule: a little less than a quarter of the information in molecule SMILES strings and about half the information in scaffold SMILES strings. Moreover, there is some possibility that the encoding strategy we used is not optimal, so this might be an overestimate of the information content.

Scaffold networks are closely related to scaffold trees and scaffold groups, inheriting many of their properties. For example, the level of scaffolds in the network directly corresponds with the number of rings in the scaffold. Likewise, there is a weak correlation between molecular weight and a scaffold's level in the network. Moreover, because the number of molecules associated with each scaffold is known, the most frequent scaffolds may be identifiable. Similarly, this vulnerability can be countered by splitting the most frequent nodes.

Networks record a many-to-many relationship between scaffold parents and children. This property arises from the way scaffolds are decomposed to identify their parents. Trees are formed by picking one ring to remove from each scaffold to determine a single parent. In contrast, networks identify multiple parents for every scaffold, each one identified by removing a different ring. By enumerating all possible decompositions of scaffolds, networks can more reliably identify important substructures that are sometimes missed by trees.

Each scaffold is decomposed into network of interconnected scaffolds, in an intricate pattern. This pattern conveys information about the scaffold, and many scaffolds produce different patterns. Encouragingly, we find that networks do not provide enough information to reverse engineer structures with any certainty. We are able to show this in two ways. First, scaffolds with very different chemical structures can yield networks with the same topology. Second, there are usually millions of very different scaffolds that can generate a single scaffold network topology. This many-to-one relationship between scaffolds and specific topologies ensures that scaffold structures cannot be reverse engineered from the network. Both these points are more deeply explored here.

Ignoring the chemical structures associated with each node in this network, we say that scaffolds belong to the same "equivalence class" if they produce scaffold networks with the same topology when decomposed (Figure 7). Our PubChem data set has 4.9 million scaffolds, which fall into about 158 000 equivalence classes, about 31 scaffolds per class. This demonstrates a high degeneracy between the network topology and chemical structure. Moreover, a key finding in this study is that scaffolds in the same equivalence class are structurally dissimilar from one another. The pairwise Tanimoto similarity between scaffolds in the same equivalence class was, on average, only 0.205. Most scaffolds (94%) correspond with one of 24,000 equivalence classes with structurally diverse scaffolds: Tanimoto similarity less than 0.5.

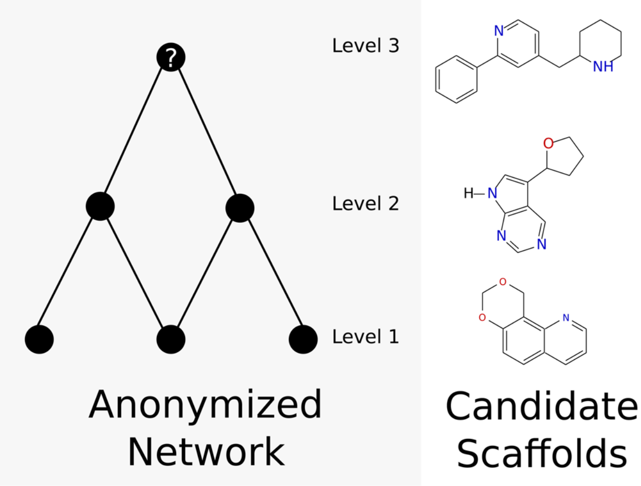


Figure 7. Scaffolds shown at the right all generate the network topology shown at the left. They all have the same number of rings and are asymmetric, but in all other ways are different. In this sense, scaffold networks convey some information about the number of rings in a molecule, but little else.

About 6% of scaffolds produce a scaffold network that can reveal the structure in some circumstances. From the prior analysis, the remaining equivalence classes fall into two cases: (1) about 82 000 are unique within our PubChem data set, with only one scaffold associated with them, and (2) about 52 000 are associated with more than one scaffold with Tanimoto similarity greater than 0.5. These two cases require further attention, because it is here that scaffold network topology could most closely correspond with chemical structure. We will see that there are many molecules (not found in our PubChem data set) that could have yielded these networks and that those in the second case correspond to highly symmetric molecules, which are not usually interesting to most drug discovery efforts. More importantly, we will show how a network can be pruned to render these scaffolds identifiable.

A scaffold network topology places constraints on the structure of a scaffold that could have generated it. For example, Figure 8 shows a scaffold network for a scaffold that must have an axis of symmetry about its central ring. With these constraints in mind, which can be complex, we could not find an exact way of enumerating all the scaffolds that could generate an arbitrary scaffold network. We could, however, compute a conservative lower bound on the number of scaffolds that could generate the network. The algorithm for computing this bound starts with an example scaffold, and detects all the symmetries contain in the scaffold. These symmetries are used to identify the number of ring and linker choices possible, and symmetry constraints on these choices. PubChem was used as a library of allowable rings (from 3 to 9 atoms) and we enumerated possible linkers (up to 5 atoms, allowing carbon, nitrogen, oxygen, and carbonyl groups). Multiplying together the number of possible rings and linkers at each position gives an estimate of the number of scaffolds in the same equivalence class with the example scaffold. This estimate is lower than the actual quantity because it does not count scaffolds with totally different ring topologies that can also yield the same scaffold network. This algorithm is discussed in greater detail in the Supporting Information.

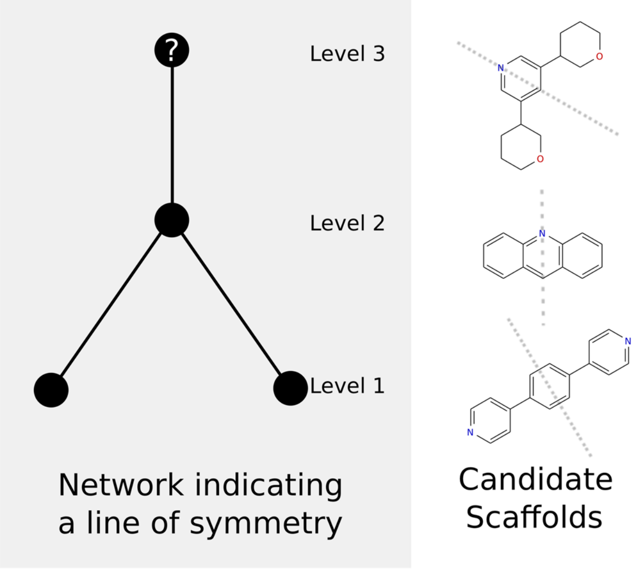


Figure 8. Scaffolds networks can impose constraints on the molecules with which they are consistent. For example (left), the scaffold network topology requires a three ring structure with an axis of symmetry about its central ring. Some examples consistent with this constraint (right), all yielding the same network, are shown.

Equivalence classes are not specific signatures for molecules because they are consistent with a very large number of possible scaffolds. For equivalence classes with either one example in PubChem or with intragroup similarity greater than 0.5, we counted the number of scaffolds that could have produced the same network. Scaffolds from these equivalence classes have, on average, eight rings. There are about 1700 possible rings and 1700 linkers to from which to choose. So a simple lower bound estimate would choose 8 rings and 7 linkers, yielding about $1700^{15} = 10^{42}$ possibilities. This estimate is not exact, because it does not take into account symmetry and asymmetry constraints and it assumes rings are connected in a linear sequence. A more accurate estimate shows that on average there are more than 10^{52} scaffolds that would have been consistent with each equivalence class.

A few equivalence classes impose a large number of symmetry constraints, and a much smaller number of scaffolds are consistent with them (Figure 9). These cases involve molecules that are not interesting in drug discovery, and could be removed from a data set before sharing. Overall, these results are just as we would expect, based on the results from information theory. Scaffold networks do not convey enough information to reliably reveal chemical structure.

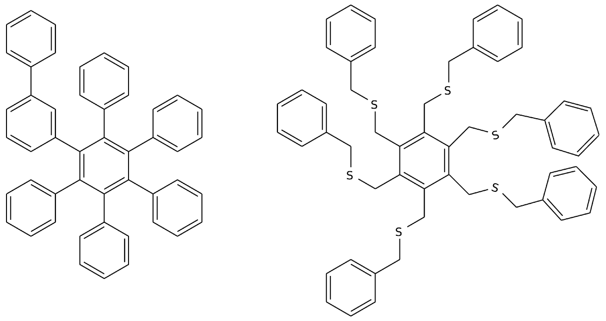


Figure 9. Example scaffolds with low diversity in the PubChem data set. We estimate that about 15282 scaffolds could produce the same networks as these scaffolds. Of particular note is the high degree of symmetry in these scaffolds, which make them much less interesting in the drug discovery and screening.

Nonetheless, in certain cases, some scaffolds' structures can be revealed if they have a unique scaffold topology. For example, let us say a company releases an anonymized scaffold network of their screening library, with assay results attached. Now imagine we know or accurately guess that this library includes molecules from a commercial provider, which makes its structures publicly available. In about 6% of cases, we can map scaffolds between the commercial provider's structures and the private company's data. This vulnerability occurs because some scaffolds are (1) in both data sets, (2) decompose in a unique scaffold network pattern, and (3) this pattern is consistently represented in both public and private data.

We can counter this attack by pruning the scaffold network to remove the nodes that are not useful for chemical analysis. The pruned network removes all scaffold-nodes from the network that are both (1) not represented by actual molecules in the database and (2) are not substructures of two or more other scaffolds in the network. Pruning these nodes ensures that no scaffolds can be identified. First, pruning the network reduces the amount of information from 8.91 bits to 1.50 bits. Second, pruning the network makes it so the same scaffold can yield different network topologies in different data sets, so it is represented differently in the public and private data. Third, pruning the network simplifies the scaffold network patterns so that very few remain unique.

This technique removes of 70% of nodes appearing in the topologies of unique scaffold networks in the PubChem data set. Of the approximately 82000 unique topologies, 11548 of them are corrupted (because they include pruned nodes) by this process. After pruning, only 1.5% of the scaffolds from the original data set are unique. Each of these 1.5% of scaffold's networks correspond to only one scaffold appears the data set, but still corresponds to a large number of potential scaffolds. Moreover, because of the pruning process, the exact scaffold network associated with each scaffold depends on the entire database. With pruning, scaffolds yield different networks in different databases, so there is no way to use a dictionary attack to reveal their structures.

■ CONCLUSIONS

Two themes unfold across this analysis. First, it does seem possible to share useful chemical information without revealing structures. Our analysis has focused on the fact that sharing strategies which reveal *relationships* among molecules do not expose structures, but they do share useful information. We emphasize that the relationships between molecules are exactly the information used for most algorithms in chemical informatics, and this is the information that is safe to share.

Second, the structures most vulnerable to exposure in these sharing strategies are the common structures and those with very high symmetry. These structures are also the least important to protect because they are usually not the most important molecules in a data set. Nonetheless, simple strategies, like node splitting or pruning, can counter these vulnerabilities.

From here, there are several important paths forward. There is still work necessary to determine the best way to split nodes in these representations so as to preserve useful information but eliminate an attacker's ability to reverse engineer scaffolds. How many nodes should be split and how many times should they be split? How should connections between nodes be distributed to the split nodes? To what extent will splitting nodes affect the accuracy of downstream analysis?

There is also a need for software to generate these representations in a secure way and to identify specific areas of chemical space that might be vulnerable when shared. We think we have identified most of the vulnerabilities in these methods, but we cannot be sure until a large community of experts review our approach in detail. As is the case with all proposals in secure sharing, the dialogue between secure methods and attempts to crack them is what ultimately establishes confidence in those methods.

Nonetheless, the same theoretical framework that demonstrated that it is not safe to share structural descriptors also provides strong evidence that sharing the relationships between molecules is secure. We hope that this paper reopens study in the secure sharing of molecules, which has remained relatively quiet for the last eight years. Adoption of these methods could help improve the efficiency of drug discovery from screening data, especially as outsourcing early discovery becomes more common. More importantly, it may enable proprietary data from chemical assays to be shared in a format that enables the analysis of hidden structures.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Algorithm for estimation of the number of scaffolds. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: swamidass@wustl.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Pathology and Immunology Department of Washington University in St. Louis for the support.

■ REFERENCES

(1) Posner, B. A.; Xi, H.; Mills, J. E. J. Enhanced HTS hit selection via a local hit rate analysis. *J. Chem. Inf. Comput. Sci.* **2009**, *49*, 2202−2210.

(2) Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **2003**, *8*, 634−647.

(3) Varin, T.; Gubler, H.; Parker, C.; Zhang, J.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound set enrichment: A novel approach to analysis of primary HTS data. *J. Chem. Inf. Comput. Sci.* **2010**, *50*, 277−279.

(4) Swamidass, S. J.; Calhoun, B. T.; Bittker, J. A.; Bodycombe, N. E.; Clemons, P. A. Enhancing the rate of scaffold discovery with diversity-oriented prioritization. *Bioinformatics* **2011**, *27*, 2271−2278.

(5) Swamidass, S. J.; Calhoun, B. T.; Bittker, J. A.; Bodycombe, N. E.; Clemons, P. A. Utility-aware screening with clique-oriented prioritization. *J. Chem. Inf. Comput. Sci.* **2011**, *52*, 29−37.

(6) Swamidass, S. J. Using economic optimization to design high-throughput screens. *Future Med. Chem.* **2013**, *5*, 9−11.

(7) Makarenkov, V.; Kevorkov, D.; Zentilli, P.; Gagarin, A.; Malo, N.; Nadon, R. HTS-Corrector: Software for the statistical analysis and correction of experimental high-throughput screening data. *Bioinformatics* **2006**, *22*, 1408.

(8) Makarenkov, V.; Zentilli, P.; Kevorkov, D.; Gagarin, A.; Malo, N.; Nadon, R. An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **2007**, *23*, 1648.

(9) Seiler, K.; George, G.; Happ, M.; Bodycombe, N.; Carrinski, H.; Norton, S.; Brudz, S.; Sullivan, J.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N.; Schreiber, S.; Clemons, P. ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **2008**, *36*, D351.

(10) Calhoun, B. T.; Browning, M. R.; Chen, B. R.; Bittker, J. A.; Swamidass, S. J. Automatically detecting workflows in pubchem. *J. Biomol. Screen.* **2012**, *17*, 1071−1079.

(11) Browning, M. R.; Calhoun, B. T.; Swamidass, S. J. Managing missing measurements in small-molecule screens. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 469−478.

(12) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M.; Waldmann, H. The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 47−58.

(13) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Comput. Sci.* **2011**, *51*, 258−266.

(14) Wassermann, A. M.; Bajorath, J. Directed R-group combination graph: A methodology to uncover structure−activity relationship patterns in a series of analogues. *J. Med. Chem.* **2012**, *55*, 1215−1226.

(15) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(16) Howells, J.; Gagliardi, D.; Malik, K. Sourcing knowledge: R&D outsourcing in U.K. pharmaceuticals. *Int. J. Technol. Manage.* **2012**, *59*, 139−161.

(17) Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Nicely, H.; Khoury, R.; Biros, M. High-throughput screening: Update on practices and success. *J. Biomol. Screen.* **2006**, *11*, 864.

(18) McGee, J. Outsourcing and contract services. *J. Biomol. Screen.* **2012**, *17*, 1379−1381.

(19) Swamidass, S. J. Mining small-molecule screens to repurpose drugs. *Briefings Bioinf.* **2011**, *12*, 327−335.

(20) Baell, J.; Holloway, G. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(21) Bradley, D. Share and share alike. *Nat. Rev. Drug Discovery* **2005**, *4*, 180−180.

(22) Masek, B. B.; Shen, L.; Smith, K. M.; Pearlman, R. S. Sharing chemical information without sharing chemical structure. *J. Chem. Inf. Comput. Sci.* **2008**, *48*, 256−261.

(23) Balaban, A. Can topological indices transmit information on properties but not on structures? *J. Comput.-Aided Mol. Des.* **2005**, *19*, 651−660.

(24) Bologa, C.; Allu, T.; Olah, M.; Kappler, M.; Oprea, T. Descriptor collision and confusion: Toward the design of descriptors to mask chemical structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 625−635.

(25) Clement, O.; Güner, O. Possibilities for transfer of relevant data without revealing structural information. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 731−738.

(26) Filimonov, D.; Poroikov, V. Why relevant chemical information cannot be exchanged without disclosing structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 705−713.

(27) Kaiser, D.; Zdrazil, B.; Ecker, G. Similarity-based descriptors (SIBAR)—A tool for safe exchange of chemical information? *J. Comput.-Aided Mol. Des.* **2005**, *19*, 687−692.

(28) Trepalin, S.; Osadchiy, N. The centroidal algorithm in molecular similarity and diversity calculations on confidential datasets. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 715−729.

(29) Tetko, I. V.; Abagyan, R.; Oprea, T. I. Surrogate data—A secure way to share corporate data. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 749−764.

(30) Karr, A. F.; Feng, J.; Lin, X.; Sanil, A. P.; Young, S. S.; Reiter, J. P. Secure analysis of distributed chemical databases without data integration. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 739−747.

(31) Faulon, J.-L.; Brown, W. M.; Martin, S. Reverse engineering chemical structures from molecular descriptors: How many solutions? *J. Comput.-Aided Mol. Des.* **2005**, *19*, 637−650.

(32) Bzip2, version 1.0.6. http://www.bzip.org (accessed July 7, 2013).

(33) James, C. A.; Weininger, D.; Delany, J. (Daylight Chemical Information Systems) Daylight Theory Manual, 2005. http://www.daylight.com/dayhtml/doc/theory/ (accessed July 7, 2013).

(34) Weininger, D. Method and apparatus for designing molecules with desired properties by evolving successive populations. U.S. Patent 5,434,796, 1995.

(35) Faulon, J.-L.; Churchwell, C. J.; Visco, D. P. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.

(36) Hohman, M.; Gregory, K.; Chibale, K.; Smith, P.; Ekins, S.; Bunin, B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discovery Today* **2009**, *14*, 261−270.

(37) Gupta, R.; Gifford, E.; Liston, T.; Waller, C.; Hohman, M.; Bunin, B.; Ekins, S. Using open-source computational tools for predicting human metabolic stability and additional ADME/TOX properties. *Drug Metab. Dispos.* **2010**, *11*, 2083−2090.

(38) Allison, M. NCATS launches drug repurposing program. *Nat. Biotechnol.* **2012**, *30*, 571−572.

(39) Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S. PubChem: Integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217−241.

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

(41) McKay, B. D. *Nauty User's Guide*, version 2.4; Computer Science Dept., Australian National University: Acton, Australia, 2007.

(42) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, No. Article 33, http://www.jcheminf.com/content/3/1/33.

(43) O'Boyle, N.; Morley, C.; Hutchison, G. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*, No. Article 5, http://journal.chemistrycentral.com/content/2/1/5.

(44) Liu, W.-H.; Sherman, A. H. Comparative analysis of the Cuthill−McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices. *SIAM J. Num. Anal.* **1976**, *13*, 198−213.

(45) Sporzynski, S. Implementation and use of high-performance parallel skyline matrix subroutines on the IBM vector facility. *Comput. Syst. Eng.* **1991**, *2*, 203−215.

(46) Silva, M. Sparse matrix storage revisited. In *Proceedings of the 2nd Conference on Computing Frontiers*; Association for Computing Machinery: New York, NY, 2005; pp 230−235.

(47) Itti, L.; Baldi, P. Bayesian surprise attracts human attention. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 547.