JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

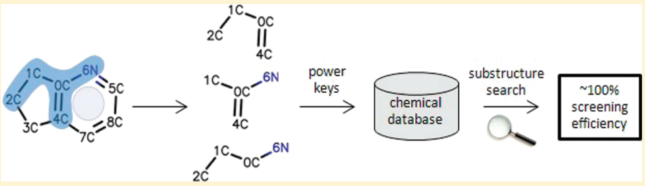
pubs.acs.org/jcim

Power Keys: A Novel Class of Topological Descriptors Based on Exhaustive Subgraph Enumeration and their Application in Substructure Searching

Pu Liu,*,⊥ Dimitris K. Agrafiotis, and Dmitrii N. Rassokhin

Johnson & Johnson Pharmaceutical Research & Development, L.L.C., Welsh & McKean Roads, Spring House, Pennsylvania 19477, United States

ABSTRACT: We present a novel class of topological molecular descriptors, which we call power keys. Power keys are computed by enumerating all possible linear, branch, and cyclic subgraphs up to a given size, encoding the connected atoms and bonds into two separate components, and recording the number of occurrences of each subgraph. We have applied these new descriptors for the screening stage of substructure searching on a relational database of about 1 million compounds using a diverse set of reference queries. The new keys can eliminate the vast majority (>99.9% on average) of nonmatching molecules within a fraction of a second. More importantly, for many of the queries the screening efficiency is 100%. A common feature was identified for the molecules for which power keys have perfect discriminative ability. This feature can be exploited to obviate the need for expensive atom-by-atom matching in situations where some ambiguity can be tolerated (fuzzy substructure searching). Other advantages over commonly used molecular keys are also discussed.



INTRODUCTION

Defining how chemical substances relate to each other has been debated long before the field of chemoinformatics even existed. Chemoinformatics has unified a broad range of computer-based methods into a coherent discipline, with molecular representation as the centerpiece. Nowadays, chemoinformatics approaches have become an integral part of research, particularly in drug discovery. For example, diversity analysis[1,2] and druglike profiling[3] can significantly enhance the value of a screening collection and the probability of finding hits that can be turned into sustainable leads; ligand-based virtual screening[4] can identify potential drug candidates from a haystack of candidate molecules without performing time-consuming and expensive wet-bench experiments or without even knowing the identity of the target; and the understanding of structure–activity relationships (SARs)[5] can give valuable insights into hit prioritization and lead optimization. The success of such systems depends critically on the methods used to represent and compare molecules.

There are many different ways to quantify chemical structures. Currently, there are thousands of descriptors available through commercial and academic software packages, which range greatly in data type, origin, sophistication, speed of calculation, and applicability to different problem domains.[6] These descriptors can be roughly classified by the dimensionality of the information used in their calculation.[7−12]

One-dimensional (1D) descriptors include parameters describing global molecular properties, such as molecular weight, molecular refractivity, number of hydrogen-bond donors/acceptors, total dipole, log P,[13] and various atom counts like Ghose and Crippen's.[14] Because they are easy to compute, these descriptors are used extensively in high-throughput processes, such as in filtering molecules for library enhancement and virtual screening.[15]

Two-dimensional (2D) descriptors are derived from connection tables and encode the topology of the molecular graph.[16] Some well-known descriptors in this class are molecular connectivity indices, κ indices, and molecular edge distances.[17,18] Structural keys, a particularly popular class, encode the presence or absence of a predefined set of structural fragments (or patterns). Structural keys are usually encoded as a binary array (fingerprint) where each bit denotes the presence or absence of the corresponding pattern. For example, public MACCS keys use an ensemble of 166 structural patterns,[19,20] whereas Barnard Chemical Information (BCI) fingerprints use more than 1000 different substructures, including augmented atoms, atom pairs, ring fusion fragments, etc.[21] Clearly, the selection of patterns has major impact on performance. Hashed fingerprints address this subjectivity and lack of generality by enumerating all possible patterns of interest from its connection table. Since the number of unique patterns can be very large, rather than assigning a unique bit to each pattern, a random number generator is employed to hash each pattern into an overlapping bit segment. Daylight[22] and Unity[23] are two such examples. Avalon fingerprints[24] is an interesting variant that incorporates both 2D and 1D information.

Closely related to fingerprints are fragment descriptors, which are constructed by enumerating all the paths present in a molecule up to a certain level.[6] For instance, multilevel neighborhoods of atoms (MNA) descriptors iteratively list the types of each atom and its neighbors. The descriptor at each successive level is obtained by concatenating the descriptor and the types of the neighboring atoms in the current level.[25] The signature molecular descriptors are a collection of atom codes, each of

Received: June 18, 2011
Published: September 28, 2011

dx.doi.org/10.1021/ci200282z

which is a canonical representation of the subgraph containing all the atoms that are at a predefined topological distance from the central atom.[26] Two more recent examples are the Molprint 2D and the extended connectivity fingerprints (ECFP).[27] The Molprint 2D algorithm encodes the atomic environment by assigning a Sybyl type to each atom, counting the occurrences of each atom type at a given number of bonds from the central atom, and encoding this information into a set of strings.[7,8] In ECFP, each atom is assigned a code based on the number of connections, the element type, the charge, and the atomic mass. The code of the central atom and its neighbors and the bonds between them is combined and hashed to produce the fingerprint of the next level.[28] Another type of fragment descriptors is involved in the information retrieval using graph kernel methods, where a molecular graph is decomposed into shortest paths,[29] subtrees,[30] cycles,[31] or other substructures from depth-first sesarch.[32] A prominent example is the successful application of graph kernels on predicting mutagenicity, toxicity and anticancer activity by counting labeled paths up to certain depth from each vertex.[32]

Going beyond 2D connectivity, three-dimensional (3D) descriptors are derived from the 3D structure of the molecule and encode features that are conformation-dependent. Many of them are based on the century-old idea of pharmacophores, that is, molecular frameworks that carry essential features responsible for a drug's biological activity. Examples of such 3D descriptors include a large variety of pharmacophore fingerprints,[33] such as CATS,[34] Similog,[35] FEPOPS,[36] TGD,[37] molecular field descriptors,[38,39] and shape-based fingerprints.[40,41] These descriptors are particularly useful for the fourth type of chemical search, namely pharmacophore search, but they are significantly more expensive to compute and often suffer from incomplete sampling of conformational space.

While the aforementioned descriptors have found extensive use in similarity searching, clustering, and QSAR, most of them were designed for flat files and are not optimized for large relational databases. More importantly, only few of them can be used for substructure searching. Substructure searching is the process of retrieving all the molecules in a chemical database that contain a specific query structure. It is widely used in chemical structure mining for detecting and analyzing crucial molecular features responsible for important physiochemical and biological properties. Although from a graph-theoretic viewpoint substructure searching is an NP-complete subgraph isomorphism problem,[42,43] it is still possible to develop systems with acceptable average time performance,[44] especially for molecular graphs which are generally characterized by low connectivity.

Since the publication of the first back-tracking algorithm by Ray and Kirsch in 1957,[45] and because of its profound importance in the chemical and pharmaceutical sciences, the development of efficient substructure search algorithms has attracted much attention in the chemoinformatics literature.[43,46] The performance of the original back-tracking algorithm was further improved by utilizing more detailed labels and other techniques.[47−49] Another class of algorithms uses partitioning to reduce the number of mappings and relaxation to iteratively refine the search.[43] Examples include Sussenguth's algorithm,[50] Figueras's set reduction algorithm,[51] Ullmann's algorithm,[52] and Von Scholley's algorithm.[53]

For chemical databases containing a large number of molecules, a linear scan of the database using this atom-to-atom matching approach would be very time-consuming and inefficient. In most chemical database management systems (DBMS), the actual search involves two distinct steps: (1) screening and (2) verification. The screening phase involves a rapid search through the database to identify molecules that can be safely excluded from further consideration because they do not contain features that are present in the query pattern.[54] For example, if the query contains a chloro-phenyl group, any molecules in the database which do not contain aromatic or cholrine atoms can easily be identified and eliminated from further consideration. The molecules that pass this screening stage are verified one-by-one using any of the aforementioned isomorphism algorithms. This molecule-by-molecule comparison is by far the rate limiting step in the entire process, so it is essential that the screening selectivity (defined as the ratio of verified hits over screening hits) be as high as possible.

Naturally, the selection of features (or keys) has a profound impact on the overall performance of the system. Keys which are not sufficiently discriminatory (e.g., the presence of a carbon atom) will yield poor selectivity. Keys are typically selected based on their statistical occurrence in large chemical databases. The more advanced screening approaches utilize hierarchical tree structures, such as atom-centered fragments and ring descriptors. This class of approaches has been widely adopted in various systems[55−64] and extended to subgraphs.[65]

Recently, an innovative algorithm designed specifically for relational databases was reported by Golovin and Henrick.[44] This method casts the subgraph isomorphism problem into a standard SQL query without necessitating external function calls. By analyzing the symmetry of a query molecule and by performing breadth-first searches, the speed of their implementation in searching relational DBMSs is quite remarkable. For a database of about 1 million compounds, a typical search can be performed in a few seconds.

Here, we describe a new class of molecular descriptors called "power keys" that are computed directly from molecular graphs and show impressive performance in substructure searching. In the following sections, we describe the algorithm used to generate them and demonstrate their ability to improve screening efficiency using a diverse set of substructure queries performed on a SQL Server relational database comprising approximately 1 million compounds. We show that power keys can eliminate the vast majority of (and in many cases 100%) nonmatching molecules within a fraction of a second. The accuracy and performance of the new method is compared to the substructure search method developed by Golovin and Henrick.[44] The screening selectivity of the power keys was also compared with the selectivity of MACCS,[19,20] CACTVS,[66] and the OpenBabel FP2 fingerprints.[67]

The power keys possess several characteristics that make them particularly suitable for substructure screening. First, they have very good screening selectivity. Moreover, for molecules with certain common features, the discriminating ability of power keys is 100% accurate for all practical purposes, which can alleviate the need for expensive atom-by-atom matching in fuzzy substructure searching. It should also be noted that power keys can be used both for substructure and similarity searching, obviating the need to maintain separate indices for each task, and thus significantly reducing storage requirements in large databases providing high-performance substructure and similarity search capabilities.

■ METHODS

Power Key Generation. Power keys are numerical codes derived from and associated with every unique subgraph of a given size that is present in the target molecule. Indeed, the

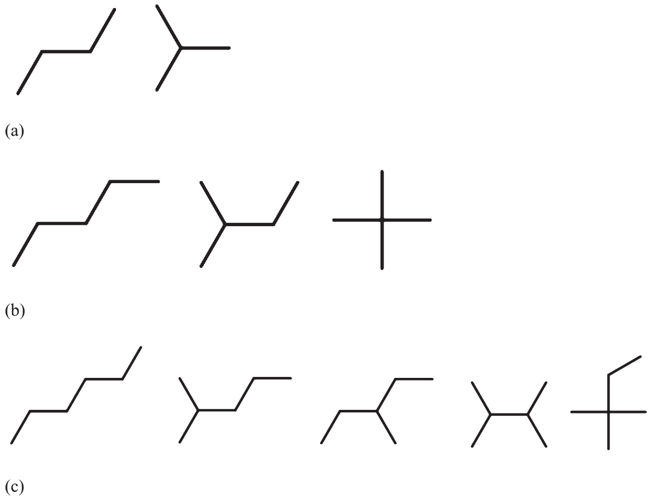


Figure 1. Possible topological configurations for a molecular subgraph with no more than four valence bonds on each atom. As shown in subgraph a−c, there are 2, 3, and 4 different configurations for molecular subgraphs containing 4, 5, and 6 atoms, respectively.

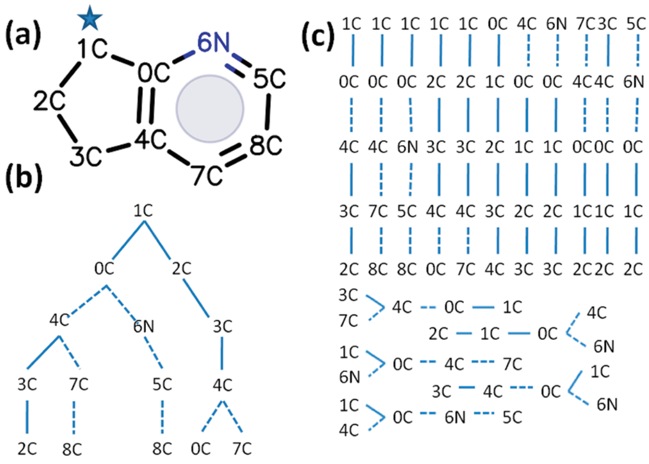

(a)
1C
6N
0C
5C
2C
4C
8C
3C
7C
(b)
1C
0C
2C
4C
6N
3C
3C
7C
5C
4C
2C
8C
8C
0C
7C
(c)
1C 1C 1C 1C 1C 0C 4C 6N 7C 3C 5C
0C 0C 0C 2C 2C 1C 0C 0C 4C 4C 6N
4C 4C 6N 3C 3C 2C 1C 1C 0C 0C 0C
3C 7C 5C 4C 4C 3C 2C 2C 1C 1C 1C
2C 8C 8C 0C 7C 4C 3C 3C 2C 2C 2C
3C 7C 4C 0C 1C
2C 1C 0C 4C 6N
1C 6N 0C 4C 7C
3C 4C 0C 1C 6N
1C 4C 0C 6N 5C

Figure 2. Schematic illustration of the first few steps in generating the power keys of size 5 for the paths originating from the atom labeled with a star. (a) Molecular graph for 2,3-cyclopentenopyridine and sequential atom numbering. (b) Subgraph consisting of all atoms up to 4 bonds away from atom 1 extracted from the molecular graph. The solid and dashed lines in the tree denote the single and aromatic bonds, respectively. (c) All the subgraphs present in c that include atom 1 anywhere on the path exhaustively enumerated. As detailed in the main text, these subgraphs are encoded as a pair of integers.

properties of a molecule are determined by its constituent atoms and the bonds connecting them. In chemoinformatics, a molecule is typically represented as a graph $G = (V_G, E_G, C(V,E))$, where V_G is the set of vertices or nodes (atoms), E_G is the set of edges (bonds), and $C(V,E)$ is the coloring scheme used for the nodes and edges, respectively. Colors are labels which distinguish different atom and bond types (e.g., atomic element and charge/hybridization state, bond order, or some other convention pertinent to the task at hand). For a typical organic molecule, the degree of an atom d, i.e., the number of atoms covalently attached to it, is usually no greater than 4.

Given a graph G, all the connected subgraphs containing a prescribed number of nodes n can be easily enumerated. As shown in Figure 1, if we consider only acyclic paths (linear and branched) and ignore atom and bond labels, there is only one possible acyclic topology for subgraphs with 1, 2, and 3 nodes, while there are 2, 3, and 5 different acyclic topologies for subgraphs of 4, 5, and 6 nodes, respectively. All the cyclic topologies of size n occurring in a particular molecule can be generated by identifying all the acyclic topologies of the same size and determining if any of the atoms are covalently bonded in the parent structure.

Table 1. Atom Types Used to Calculate Atom Weights in Equation 1

atom type	weight	atom type	weight
H	0	Al	11
He[a]	1	Si	12
Li	2	P	13
Be	3	S	14
B	4	Cl	15
C	5	K	16
N	6	Ca	17
O	7	R[b]	18
F	8	Br	19
Na	9	I	20
Mg	10		

[a] He, Ne, Ar, Kr, Xe. [b] The rest of elements in the periodic table.

Table 2. Bond Types Used to Calculate Atomic Weights in Equation 1

bond type	weight
single	0.00001
double	0.00002
triple	0.00004
aromatic	0.00008

Let the symbol a be a node (atom) in the molecular graph G, n the number of nodes in a subgraph of G, $g(n, a)$ a subgraph of G containing node a and $(n - 1)$ other nodes connected to it, and $G(n, a)$ the set of all subgraphs $g(n, a)$ for a given a and n. The power keys for node a can be constructed from the subgraphs $G(n, a)$ by the following procedure, which is partially illustrated in Figure 2 for 2,3-cyclopentenopyridine:

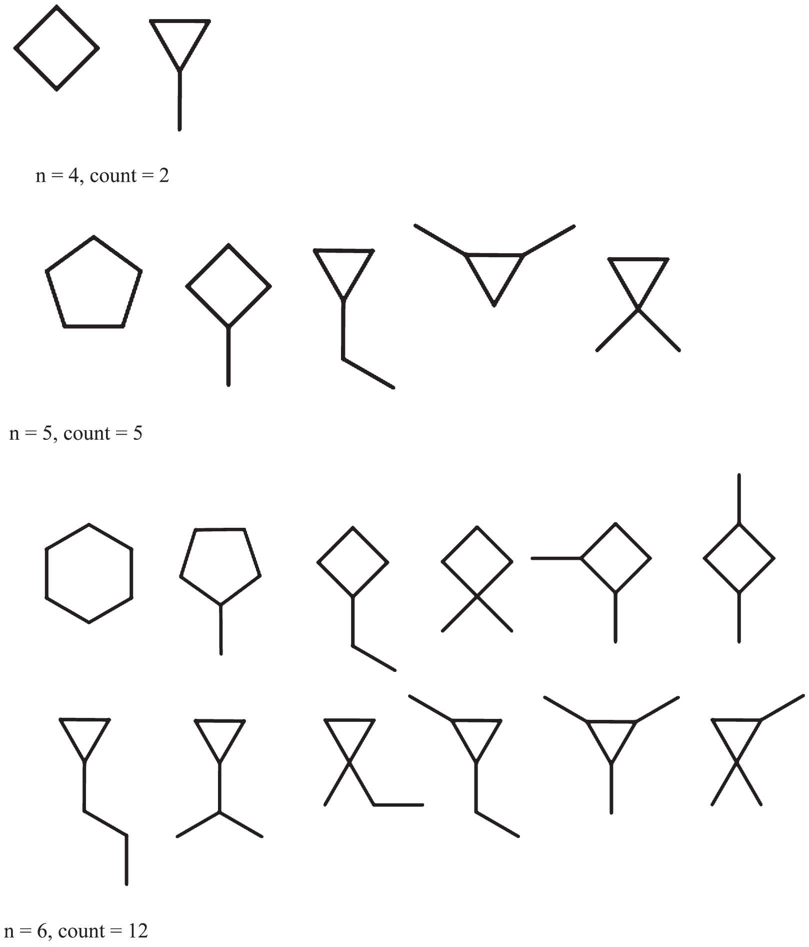
(a) Each atom in the molecule is numbered sequentially, with atom a assigned index 1 (Figure 2a).

(b) All the acyclic subgraphs emanating from atom a and containing up to $(n - 1)$ bonds are exhaustively enumerated by traversing the molecular graph G in a recursive, depth-first manner starting from atom a, and avoiding any already visited atoms along the way to avoid ring closures (Figure 2b).

(c) This process is repeated for every atom in the molecule, and all the subgraphs containing atom a and $(n - 1)$ other connected atoms are stored in a list $(G(n, a))$. Note that in the subgraphs $G(n, a)$, a can appear anywhere on the path and need not be a terminal node (Figure 2c).

(d) For each subgraph from step c, the numerical labels are replaced with the atom types listed in Table 1. Each atom is assigned a weight using the following formula:

$$w_a(a) = w_N(a) + \sum_{i=1}^{d} w_E(b_i) \tag{1}$$



n = 4, count = 2

n = 5, count = 5

n = 6, count = 12

Figure 3. Possible cyclic topologies for a molecular subgraph with one ring closure and no more than 4 valence bonds on each atom. There are 2, 5, and 12 different cyclic configurations for subgraphs containing 4, 5, and 6 atoms, respectively.

where $w_N(a)$ is the weight associated with atom a as listed in Table 1, b_i is the ith bond connected to atom a, d is the number of atoms attached to a, and $w_E(b_i)$ is the weight associated with bond b_i as listed in Table 2. These weights are used to distinguish different bonding environments for each atom and facilitate the canonicalization of the atom sequence in the next step. Thus, at the end of this step, each subgraph is associated with a weight array of size n.

In order to remove any ambiguity introduced by different ways of atom ordering and subgraph traversal, the weight arrays need to be canonicalized. The canonicalization algorithm is slightly different for different topologies:

For linear subgraphs, such as the first one in each panel in Figure 1, the nodes are sequentially numbered (ordered) from one end to the other in a way that maximizes the expression:

$$T_p \operatorname{pow}(A, n) + \sum_{i=1}^{n} w_a(a_i) \operatorname{pow}(A, i-1) \qquad (2)$$

where n is the number of nodes in the subgraph, A is the number of different atom types in Table 1 (i.e., 21 in the present work), a_i is the ith node in the linear sequence, $w_a(a_i)$ are the weights computed in step d above, pow is the mathematical function power, and T_p is a topological constant uniquely assigned to each possible acyclic and cyclic topology for subgraphs containing up to 6 atoms. In the present work, only linear, branched, and simple cyclic topologies with one ring closure were considered. As illustrated in Figure 3, there are 1, 2, 5, and 12 one-ring topologies for subgraphs with 3, 4, 5, and 6 nodes, respectively. While cyclic topologies with more than one ring and/or more than four atoms connected to a central atom are in theory possible, such substructures occur very rarely and their inclusion would not appreciably improve screening efficiency and substructure search performance (nor would their omission lead to any false negatives in the search—vide infra). It is worth emphasizing that for a single cyclic topology we generate two keys: one for the cyclic path and one for its acyclic counterpart. The former is a superstructure of the latter.

For nonlinear topologies, the atom with the largest degree d is identified and assigned index 1. Starting from that atom, the longest chain in the subgraph is found and the atoms are numbered sequentially. The remaining atoms in the side chains are numbered in a way that maximizes the expression in eq 2. If there are multiple atoms with the same largest degree, the process is repeated for each one, and the ordering that maximizes the expression in eq 2 is selected.

For each canonicalized subgraph $g(n, a)$, two integer keys are computed: one encoding the atoms and one encoding the bonds.

The atom key, $K_{\text{atom}}(g(n, a))$, is calculated using the formula:

$$K_{\text{atom}}(g(n,a)) = T_p \operatorname{pow}(A,n) + \sum_{i=1}^{n} w_N(a_i) \operatorname{pow}(A, i-1) \tag{3}$$

where all the symbols have the same meaning as described above. The bond key, $K_{\text{bond}}(g(n, a))$, is encoded using a similar formula:

$$K_{\text{bond}}(g(n,a)) = \sum_{i=1}^{n_b} w_b(b_i) \operatorname{pow}(B, i-1) \tag{4}$$

where n_b is the number of bonds in the subgraph, B is equal to 16, and $w_b(b_i)$ is the weight of bond b_i, set to 1, 2, 4, and 8 for single, double, triple, and aromatic bonds, respectively. Therefore, each subgraph $g(n, a)$ can be represented as a key pair $K(g(n, a)) = (K_{\text{atom}}(g(n, a)), K_{\text{bond}}(g(n, a)))$, hereafter referred to as a power key. Note that the encoding process does not involve any hashing—each unique subgraph is assigned a unique atom and bond key. This one-to-one correspondence makes it possible to reconstruct the subgraphs from the atom and bond keys and, more importantly, makes them ideally suited for substructure searching (vide infra).

The process above yields a set of power keys (key pairs),[68] each of which encodes a unique subgraph in the target molecule. Because each subgraph may occur multiple times in the same molecule, the power keys are stored in a dictionary along with their number of occurrences c_i. Thus, each entry in the dictionary is a triplet of integers: the atom key, the bond key, and the count of occurrences of that particular atom and bond key combination. If each unique power key is considered as a basis vector in chemical space, a molecule can be represented as a point in that space, whose coordinates are given by the counts c_i.

Substructure Searching. Power keys can be used to index a chemical database and improve screening efficiency during substructure searching. The process involves three steps. During database construction (indexing), the power keys of each molecule in the database are computed and stored into a dedicated table along with their corresponding counts. During an actual search, the power keys of the query pattern are generated and compared against the power keys of the molecules in the database using standard relational database machinery. For a molecule to be a potential hit (screening hit), it must contain all the power keys that are present in the query, and with at least the same number of counts. More formally, if $\{K_i(m)\}$ and $\{c_i(m)\}$ represent the set of power keys and their corresponding counts in the database molecule, and $\{K_i(q)\}$ and $\{c_i(q)\}$, the set of power keys and their corresponding counts in the query, a molecule is a potential hit if and only if:

$$\forall\ K_i(q) \in \{K_i(q)\} \rightarrow \exists\ K_j(m) \in \{K_j(m)\} : \\ K_j(m) = K_i(q) \text{ and } c_j(m) \geq c_i(q) \tag{5}$$

All the molecules that satisfy condition 5 are examined one-by-one using rigorous atom-to-atom matching to identify the true positives.

Implementation. All calculations were performed on an IBM Thinkpad T61 laptop computer equipped with a single dual-core 2 GHz mobile Intel processor and 1.96 GB 667 MHz DRAM. The core algorithms (structure parsing, molecular perception, generation of power keys, graph isomorphism) were implemented in C++ and are part of the DirectedDiversity software suite,[69] which is the foundation of the Third Dimension Explorer (3DX) and ABCD informatics systems.[68]

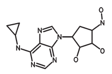
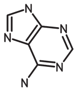
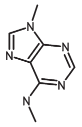
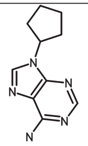
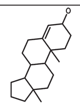
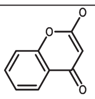
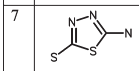
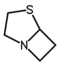
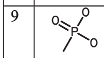
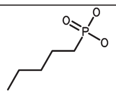
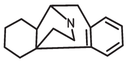
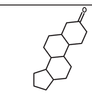
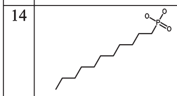
These classes are exposed to underlying Microsoft SQL Server using .Net wrappers. To allow the specification of chemical queries and their combination with other predicates, the SQL language and database query engine were extended to perform power key generation from the supplied pattern(s), subsequent screening against the database molecules based on eq 5, and the atom-by-atom matching, whenever necessary.

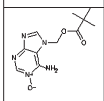
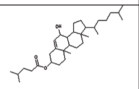
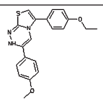
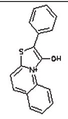
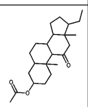
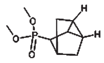
Of course, the actual database management systems include more elaborate schemes to handle insertions, deletions, and modifications of chemical records. A command line utility has also been developed that allows the indexing and searching of flat structure files provided in one of the standard formats (e.g., SDF or SMILES). For path lengths up to six, a power key table is constructed for the entire collection, where each row records a unique feature with the molecule index, power keys, and their corresponding counts. Given a query pattern, all the subgraphs, linear, cyclic, or branched with up to six atoms are exhaustively enumerated and used as filters to screen out the unmatched candidates using eq 5. The remaining molecules are passed to the atom-by-atom exact matching engine. A typical query is of the form "*select * from chem.Compound where chem.IsStrucutreMatch(EncodedMolecule, @query) <> 0 and Id in (select Id from chem.Screen('chem.Compound', 'EncodedMolecule', @query)*", where *chem.Screen()* does the fingerprint-based screening, *chem.IsStructureMatch()* does the subgraph isomorphism-based verification, *chem.Compound* is the compound table to search against, *EncodedMolecule* is an internal binary molecule format that accelerates exact matching, and *@query* is the query pattern.

RESULTS

To validate our algorithm and assess its performance, we performed a direct comparison with the method of Golovin and Henrick.[44] These authors generously provided us with the chemical database that was used in their original work,[44] which was compiled from the Open NCI database and a few other public sources and contained 931 007 molecules. Removal of stereochemically redundant SMILES and structures which could not be processed by our internal software reduced the database size to 907 558 unique molecules. When processed by our power key generation algorithm, these molecules resulted in 39 235 unique k_{atom} and 484 unique k_{bond} keys for subgraphs of size 6, and 9675 unique k_{atom} and 176 unique k_{bond} keys for subgraphs of size 5. On average, each molecule produced 41 unique keys for subgraphs of size 5 and 79 unique keys for subgraphs of size 6, respectively. The power keys and their respective counts were recorded for each molecule and uploaded into a Microsoft SQL Server 2008 database instance installed on a laptop computer (see the Implementation section above). We also compared the screening selectivity of the power keys with three other structure keys commonly used for substructure searching, the 166-bit MACCS keys[19,20] the 1024-bit path-based FP2 keys as implemented in the OpenBabel toolkit,[67] and the CACTVS keys used in PubChem and computed using Xemistry's CACTVS toolkit,[66] courtesy of Wolf-Dietrich Ihlenfeldt. To simplify the comparison, we used the same 14 query molecules that were used in the original publication by Golovin and Henrick.[44] Our effort was facilitated greatly by the availability of their search engine as a publically available web service (http://www.ebi.ac.uk/msd-srv/chemsearch). Five parallel and independent searches were

Table 3. Search/Screening Results for a Diverse Set of Query Molecules in a Chemical Database Containing ~1 Million Molecules

	Query	GH	PK	MACCS[1]	CACTVS	FP2
1		1	1	162	1	1
2		2722	2727[2]	21488[3]	13045	2777
3		534	534	17153	12785	557[3]
4		84	84	11823[3]	430	178
5		73	79[4]	6987	419	691
6		16	16	15582	11261	15[3]
7		1071	1071	3785[3]	5904	1071
8		148	148	154	149	140[3]
9		3045	3045	6221	8982	3090
10		299	299	6221	6854	303
11		150	150	1566	178	545
12		28371	29167[5]	76328[3]	28974[3]	37793[3]
13		462	1080[4]	160660	2872	12133
14		72	97[4]	6221	2118	159

Molecule a	Molecule b	Molecule c	Molecule d	Molecule e	Molecule f
					

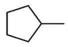
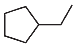
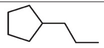
[1] In order to make the comparison valid, we had to exclude 36 MACCS fingerprint bits corresponding to the patterns with explicitly specified hydrogen atom counts: 28, 34, 43, 53, 54, 68, 69, 74, 82, 84, 86, 90, 91, 93, 100, 104, 108, 109, 111, 114, 115, 116, 118, 128, 129, 131, 132, 138, 139, 141, 147, 149, 151, 153, 155, 160. [2] The difference comes from 6 SMILES which cannot be processed by our in-house tools, and 11 additional hits in our results (such as molecule a), which may be due to slightly different definition of aromatics. This can be considered a no-false-positives case. [3] A small number of false negatives were observed (15 in the worse case of MACCS keys with pattern 2). This has been traced to a slightly different perception of aromaticity between different toolkits and does not bias the screening efficiency comparison. [4] False positives, such as molecules b, e, and f for queries 5, 13, and 14, respectively, make exact atom-by-atom matching necessary. [5] Some molecules, such as molecule d, only appear in the PK results while others, such as molecule c, are only found in the GH results.

conducted for each query molecule. The first was against the Golovin and Henrick web service, the second against our own SQL Server database indexed with the power keys and their associated counts, and the remaining three were performed by binary matching of the fingerprints computed for the query molecules against the fingerprints computed for the search set molecules. The results of the first search represent the true hits, since the process involved exact atom-by-atom matching. The results of the other searches did not involve any verification and, thus, represent the hits obtained purely from the screening step. The results are summarized in Table 3. In the remaining discussion, we will use the abbreviation GH to refer to the hit lists associated with the Golovin and Henrick method, and PK, MACCS, FP2 and CACTVS to refer to the hit lists obtained from screening with the power keys and the MACCS, FP2, and CACTVS fingerprints, respectively. Although the set of molecules used in the power key and fingerprint screening tests contained only 907 558 out of the 931 007 molecules that were present in the GH database, the excess molecules were carefully tagged and were not taken into account in the molecule-by-molecule comparison.

A good screening method is one that does not produce any false negatives and minimizes the number of false positives. The higher the number of false positives, the more time is spent on expensive atom-by-atom matching. Screening using our novel power keys required only a fraction of a second of a laptop computer and, as seen in Table 3, yielded almost identical results to Gorovin and Henrick's method. For most of the queries, the selectivity of the PK screening method was 100%, yielding identical hit lists to the GH web service (i.e., zero false negatives and zero false positives).

Closer examination of the remaining cases leads to some interesting insights. For molecule 2, our method produced five more hits than GH. After careful analysis, this difference was pinpointed to 6 SMILES which could not be processed by our toolkit and 11 additional true positives in the PK search which were missed by GH (e.g., molecule a in Table 3). We conjecture that these additional hits may come from a slightly different

Table 4. Search Results for Three Additional Query Molecules in the Same Database of ∼1 Million Molecules[3]

	Query Molecule	GH	PK	MACCS[1]	CACTVS	FP2	PK Hits Missed by GH	False Positives in PK Hits
15		26141	26160	442230	40141[2]	36260	19 (g, h)	0
16		21319	23931	442230	38955[2]	29016[2]	13 (g, i)	2599 (j)
17		16695	23439	442230	33219[2]	29016	11 (g, k)	6733 (l)

[1] See Table 3, note 1. [2] See Table 3, note 3. [3] Power search identifies extra true positives for each query molecule that were missed by the GH method. For the queries containing homogeneous keys (16 and 17), PK returns a substantial number of false positives. Representative structures for both true and false positives are listed in parentheses and shown in the secondary table underneath the main one.

definition of aromatics, a common compatibility problem between different chemoinformatics toolkits.[70] Equally importantly, there were some false positives in our search results for query molecules 5, 13, and 14. These molecules have one interesting feature in common, in that they contain paths where all the atoms and bonds are of the same type (carbon atoms connected by single bonds). Indeed, if we define a homogeneous path of size n ($n = 7$ in the present study) as a path containing n identical atoms and $(n - 1)$ identical bonds, then all these three molecules have homogeneous path of size 7 as a common feature.

To empirically verify whether this observation is indeed responsible for the reduced screening efficiency, we used three additional molecules (15, 16, and 17) as queries and summarized the results in Table 4. Molecule 15 does not contain any homogeneous paths, while molecules 16 and 17 do. For all three molecules, our method identified some additional true positives, molecule g and h, g and i, and g and k for queries 15, 16, and 17, respectively. In these additional true hits that were missed by the GH search the query patterns are embedded in multiple rings. For molecules 16 and 17 which contain homogeneous paths, our method produced many false positives (Table 4 lists one such representative structure for query 16 and 17). These results are consistent with the observation made for molecules 5, 11, 13, and 14 and provide further evidence that the presence of homogeneous paths in the query adversely impacts the screening efficiency of the power keys. For such molecules, screening performance can be further improved by employing additional keys that better capture this specific feature or encode some other properties.

Again, for molecule 15, which has no homogeneous paths, every screening hit identified by the power search was confirmed as a true positive, and the atom-by-atom matching is not necessary. This is not surprising since the encoding of all the linear, branched, and cyclic paths provides a local 2D holographic picture of the molecular graph. Although it is theoretically possible that a screening hit may not be a true hit, that probability is extremely small if there are no homogeneous keys present in the molecule, at least for the known chemical universe of ∼70 million compounds. Even when there are false positives, these molecules will most likely contain substructures that are very similar to the query pattern. Therefore, power screening may be used without verification (exact graph matching) when 100% accuracy/confidence is not required, or when a handful of false positives can be tolerated (or identified in some other way). As demonstrated in Tables 4 and 5, this kind of "fuzzy" substructure search would be immensely more efficient compared to conventional approaches, particularly when applied on a massive and repetitive scale (batch processing of large numbers of patterns on very large databases).

Table 5

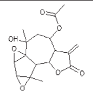
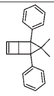
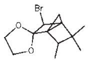
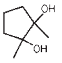
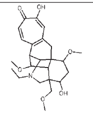
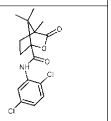
Molecule g	Molecule h	Molecule i	Molecule j	Molecule k	Molecule l
					

Table 3 also contains the results of screening based on MACCS, FP2, and CACTVS fingerprints. In order to make the comparison with the MACCS-based screening valid, we had to modify the definitions of the key generation rules used in the MACCS key computation to exclude the bits corresponding to the patterns containing explicit hydrogen atom counts, since our search patterns allow for non-hydrogen substitutions on all atoms where such substitutions are possible. An example of such an excluded rule would be rule no. 160 defined as follows: ('[CH3]',0), which means that the bit no. 160 will be set if the structure or search pattern contains at least one aliphatic carbon atom with exactly three attached hydrogen atoms. The CACTVS fingerprints for the search patterns were computed in the "SMARTS" mode, which disregards the number of attached hydrogen atoms, unless explicitly specified in the SMARTS string. For most of the tested search patterns, the screening selectivity of power keys was much better, resulting in a significantly smaller number of false positives. Only the FP2 fingerprints, which are path-based, approached power keys in their screening performance.

■ DISCUSSION

Power keys is a novel class of topological molecular descriptors derived by exhaustively enumerating, canonicalizing, and uniquely encoding all possible subgraphs up to a certain length. In this work, we have demonstrated their utility in substructure searching, where they can be used to rapidly screen a chemical database to filter out structures that could not possibly match the query of interest and, thus, minimize the number of molecules that need to be verified by expensive atom-to-atom matching.

Compared to many other molecular keys and nonhashed fingerprints, the current method does not require a predefined substructure dictionary, which removes the arbitrary element of selecting what features to use. The current method exhaustively identifies all relevant features and is therefore suitable for all types of queries and chemical databases of organic molecules. Since our method enumerates and canonicalizes all possible subgraphs of a certain size, it does not require complex symmetry handling and the resulting keys preserve the information of *tertiary* and *quaternary centers*, which is often missing for path-based molecular keys.

Unlike hashed binary fingerprints, where each feature is hashed into a segment of bits, power keys do not involve any hashing or folding. This special property, along with the fact that power keys encode not just the presence or absence of a particular path but also its count, eliminates the possibility of collisions that is inherent in the hashing process, and thus avoids information loss. Moreover, the deterministic nature of power keys makes the recovery of the actual paths from which the constituent integer components were derived straightforward.

Currently, only the atomic element is utilized as a label for encoding and canonicalizing the nodes in the paths. Although it is possible to introduce more complicated atom typing schemes (such as those employed in Molprint2D or ECFP) that could increase their discriminatory ability in other data mining applications, the resulting keys would not be suitable for conventional applications of substructure searching.

Just like many other descriptors of this kind, power keys lend themselves to a broad range of chemoinformatics and modeling applications, including similarity searching, clustering, diversity profiling, virtual screening, and QSAR. For instance, the MinMax or even the MinMax Tversky measures[32,71,72] can be employed to quantify the similarity between molecules with nonbinary fingerprints. Recently, it was further demonstrated that graphics processing units (GPU) are particularly useful in handling the relatively bulky count-based fingerprints, allowing similarity searches on the entire PubChem collection (~32 million compounds) to be completed within 1−2 s with a single GPU card.[73] Their utility of our power keys in these contexts is currently under investigation and will be presented in future publications.

AUTHOR INFORMATION

Corresponding Author

*E-mail: puliu45@gmail.com.

Present Addresses

[⊥]Lincoln Financial Group, 2005 Market Street, Philadelphia, Pennsylvania 19103, United States

ACKNOWLEDGMENT

The authors thank Dr. Golovin, Dr. Henrick and Dr. Ihlenfeldt for their generous help and insightful communications on the search results.

REFERENCES

(1) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841–51.

(2) Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries. *J. Comput. Aid. Mol. Des.* **2002**, *16*, 335–56.

(3) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model* **2010**, *50*, 470–9.

(4) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model* **2006**, 46.

(5) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

(6) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim; New York, 2002.

(7) Bender, A.; Mussa, H. Y.; Glen, R. C. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–8.

(8) Bender, A.; Mussa, H. Y.; Glen, R. C. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–18.

(9) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–33.

(10) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–53.

(11) Li, Q.; Bender, A.; Pei, J.; Lai, L. Larg A Descriptor Set and a Probabilistic Kernel-Based Classifier Significantly Improve Druglikeness Classification. *J. Chem. Inf. Model* **2007**, *47*, 1776–86.

(12) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–96.

(13) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–102.

(14) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. I. Partition coefficients as a measure. *J. Comput. Chem.* **1986**, *7*, 565–77.

(15) Dearden, J. C. In silico prediction of ADMET properties: how far have we come? *Expert. Opin. Drug. Metabol. Toxicol.* **2007**, *3*, 635–9.

(16) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573–88.

(17) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–45.

(18) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76–82.

(19) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–8.

(20) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *46*, 1273–80.

(21) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–2.

(22) James, C. A.; Weininger, D. *Daylight theory manual*; Daylight Chemical Information Systems Inc.: Irvine, CA, 2007.

(23) Software and documentation available from Tripos Associates, St Louis, MO. E-mail: support@tripos.com.

(24) Gedeck, P.; Rohde, B.; Bartels, C. QSAR—how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model* **2006**, *46*, 1924–36.

(25) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666–70.

(26) Faulon, J.-L.; Donald P. Visco, J.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–20.

(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–54.

(28) Accelrys Inc. *Pipeline Pilot*, http://accelrys.com/products/scitegic/index.html: San Diego, CA, 2009.

(29) Borgwardt, K. M. Shortest-path Kernels on Graphs. In *Prof. Intl. Conf. Data Mining*, The Fifth IEEE International Conference on Data Mining, Houston, Texas, USA, November 27−30, 2005; pp 74−81.

(30) Vert, J. P. Tree A Kernel to Analyze Hylogenetic Profiles. *Bioinformatics* **2002**, *18*, S276–284.

(31) Horvath, T.; Gartner, T.; Wrobel, S. Cyclic Pattern Kernels for Predictive Graph Mining. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, August 22−25, 2004; pp 158−167.

(32) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.

(33) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.

(34) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.

(35) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.

(36) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144–59.

(37) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.

(38) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity. *J. Chem. Inf. Model* **2006**, *46*, 665–676.

(39) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.

(40) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model* **2005**, *45*, 673–684.

(41) Wilson, J. A.; Bender, A.; Kaya, T.; Clemons, P. A. Alpha Shapes Applied to Molecular Shape Characterization Exhibit Novel Properties Compared to Established Shape Descriptors. *J. Chem. Inf. Model* **2009**, *49*, 2231–2241.

(42) Willett, P. A review of chemical structure retrieval systems. *J. Chemom.* **1987**, *1*, 139–155.

(43) Barnard, J. M. Substructure searching methods: old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.

(44) Golovin, A.; Henrick, K. Chemical Substructure Search in SQL. *J. Chem. Inf. Model* **2009**, *49*, 22–27.

(45) Ray, L. C.; Kirsch, R. A. Finding Chemical Records by Digital Computers. *Science* **1957**, *126*, 814–819.

(46) Attias, R.; Dubois, J.-E. Substructure Systems: Concepts and Classifications. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 2–7.

(47) Xu, J.; Zhang, M. New HBA Algorithm for Structural Match and Applications. *Tetrahedron Comput. Methodol.* **1989**, 2.

(48) Dengler, A.; Ugi, I. A Central Atom Based Algorithm and Computer Program for Substructure Search. *Comput. Chem.* **1991**, *15*, 103–107.

(49) Xu, J. GMA: A Generic Match Algorithm for Structural Homomorphism, Isomorphism, and Maximal Common Substructure Match and Its Applications. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 25–34.

(50) Sussenguth, E. H., Jr. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36–43.

(51) Figueras, J. Substructure Search by Set Reduction. *J. Chem. Doc.* **1972**, *12*, 237–244.

(52) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31–42.

(53) Von Scholley, A. A Relaxation Algorithm for Generic Chemical Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235–241.

(54) Lynch, M. F. Screening large chemical files. In *Chemical information systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, 1974; pp 177−194.

(55) Wiswesser, W. J. Historic Development of Chemical Notations. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258–263.

(56) Thomson, L. H.; Hyde, E.; Matthews, F. W. Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation. *J. Chem. Doc.* **1967**, *7*, 204–209.

(57) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

(58) Bremser, W. Hose - a novel substructure code. *Anal. Chem. Acta* **1978**, *103*, 355–365.

(59) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability.1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–155.

(60) Shenton, K.; Norton, P.; Fearns, E. A. Generic Searching of Patent Information. In *Chemical Structures - the International Language of Chemistry*; Warr, W. A., Ed.; Springer: Berlin, 1988.

(61) Feldmann, R. J.; Milne, G. W. A.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. An interactive substructure search system. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157–163.

(62) Bruck, P.; Nagy, M. Z.; Kozics, S. Substructure search on hierarchical tree. *Proceedings of the 11th International Online Information Meeting*, London, Dec 8−10, 1987; Learned Information: Oxford; London, 1987; pp 41−43.

(63) Hicks, M. G.; Jochum, C. Substructure search systems. 1. Performance comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 substructure search systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191–199.

(64) Ozawa, K.; Yasuda, T.; Fujita, S. Substructure Search with Tree-Structured Data. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 688–695.

(65) Pavlov, D.; Shturts, I. Chemical substructure search screening with fingerprints built with subgraph enumeration. *Rev. Adv. Mater. Sci.* **2009**, *20*, 37–41.

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

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

(68) Agrafiotis, D. K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E. P.; Konstant, P.; Leung, A.; Lobanov, V. S.; Marichal, P.; Martin, D.; Rassokhin, D. N.; Shemanarev, M.; Skalkin, A.; Stong, J.; Tabruyn, T.; Vermeiren, M.; Wan, J.; Xu, X. Y.; Yao, X. Advanced Biological and Chemical Discovery (ABCD): Centralizing Discovery Knowledge in an Inherently Decentralized World. *J. Chem. Inf. Model* **2007**, *47*, 1999–2014.

(69) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869–15872.

(70) Golovin, A.; Henrick, K. private communication; 2009.

(71) Swamidass, S. J.; Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J. Chem. Inf. Model* **2007**, *47*, 302–317.

(72) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for Small Molecules and the Prediction of Mutagenicity, Toxicity and Anti-cancer Activity. *Bioinformatics* **2005**, *21*, i359–i368.

(73) Liu, P.; Agrafiotis, D. K.; Rassokhin, D. N.; Yang, E. Accelerating Chemical Database Searching Using Graphics Processing Units. *J. Chem. Inf. Model* **2011**, *51*, 1807–1816.