J Math Chem (2015) 53:1634–1648
DOI 10.1007/s10910-015-0510-x

ORIGINAL PAPER

Enumerating and indexing many-body intramolecular interactions: a graph theoretic approach

Robert Penfold[1] · **Peter J. Wilde**[1]

Received: 22 December 2014 / Accepted: 1 May 2015 / Published online: 19 May 2015

Abstract The central idea observes a recursive mapping of n-body intramolecular interactions to $(n + 1)$-body terms that is consistent with the molecular topology. Iterative application of the line graph transformation is identified as a natural and elegant tool to accomplish the recursion. The procedure readily generalizes to arbitrary n-body potentials. In particular, the method yields a complete characterization of 4-body interactions. The hierarchical structure of atomic index lists for each interaction order n is compactly expressed as a directed acyclic graph. A pseudo-code description of the generating algorithm is given. With suitable data structures (e.g., edge lists or adjacency matrices), automatic enumeration and indexing of n-body interactions can be implemented straightforwardly to handle large bio-molecular systems. Explicit examples are discussed, including a chemically relevant effective potential model of taurocholate bile salt.

Keywords Computer simulation · Molecular modelling · Graph theory

Mathematics Subject Classification 82-08 · 05A15 · 94C15

1 Introduction

The implementation of computer code for realistically simulating the configurations and motion of molecular objects requires modelling of many-body through-bond inter-

In memoriam: Brian P. Hills.

✉ Robert Penfold
robert.penfold@ifr.ac.uk

[1] Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

actions. In turn, it is necessary to identify the participating atoms in each interaction. This paper presents a novel and exhaustive enumeration procedure exploiting the line graph transformation of the graph that encodes the molecular structure. In principle, by virtue of the recursive nature of the algorithm, straightforward extension to arbitrarily high order interactions is possible.

Application of graph theoretic methods to the general study of molecular structure, and to equilibrium statistical mechanics in particular, are not new. Significant examples include the development of molecular branching rules [1], the enumeration of isomers and the definition of topological indexes [2], as well as the analysis of discrete lattice models [3] and Mayer's cluster decomposition of the 2-body configuration integral [4].

Modelling and theory distinguish between simple materials, comprised of weakly interacting elementary units, and complex materials including objects with internal structure characterized by relatively strong coupling and typically mimicking the covalent architecture of molecular species [5]. Furthermore, intramolecular interactions can be treated either quantum mechanically (Car–Parrinello method [6] with density functional theory of electronic structure [7]) or by a classical effective potential obtained through some more or less ad hoc coarse-graining procedure [8,9]. In molecular dynamics or equilibrium Monte Carlo simulations, for example, the total intramolecular potential energy is typically decomposed as [10]

$$\mathcal{U}_{\text{intra}}\left(\{\mathbf{r}_s\}\right) = \underbrace{\sum_{\text{bonds}} \mathcal{U}_2(\mathbf{r}_i,\ \mathbf{r}_j)}_{\text{topology}} + \underbrace{\sum_{\text{bends}} \mathcal{U}_3(\mathbf{r}_i,\ \mathbf{r}_j,\ \mathbf{r}_k) + \sum_{\text{torsions}} \mathcal{U}_4(\mathbf{r}_i,\ \mathbf{r}_j,\ \mathbf{r}_k,\ \mathbf{r}_l) + \cdots}_{\text{flexibility}}, \tag{1}$$

where the sums range over suitable n-body potentials $\mathcal{U}_n$ that depend on the spatial configuration $\{\mathbf{r}_s\}$ of the constituent atoms. Similar expansions are developed for non-bonded forces too, and may also include 1-body coupling to an external field, but the indiscriminate character of these through-space interactions usually means that the identification of participating atoms is determined by a simple range parameter. For intramolecular interactions, however, the enumeration and indexing of atoms involving in each sum of (1) is subject to the constraints of molecular topology. Indeed, the 2-body bond interactions $\mathcal{U}_2$ *define* the molecular framework, while the higher order terms in (1) serve to model the more or less restricted molecular flexibility associated with bond hybridization or electronic delocalisation (e.g., aromaticity and resonance structures). The selection of these higher order potentials is often based on chemical intuition and are typically supplied to simulation software as user defined input. Mature and widely used simulation packages (e.g., GROMACS, LAMMPS, NAMD, etc.) invariably support highly optimized force fields (e.g., CHARMM, AMBER, OPLS, MMFF, etc.) that faithfully represent detailed atomistic structures. An alternative approach is to input only the molecular topology, then systematically generate all possible many-body index lists from this information and invite the user to select non-

zero force constants and appropriate functional forms for the required potentials. This paradigm is more natural for the implementation of coarse-grained models derived by thermodynamic considerations (e.g., MARTINI [11]). To facilitate this semi-automatic procedure, a graph theoretic construction is developed here to exhaustively enumerate and index arbitrary n-body intramolecular interactions starting from the description of 2-body adjacency.

The central idea in this work observes the correspondence between the hierarchy of n-body intramolecular interactions and iteration of the line graph [12] transformation $L(G)$ on a connected "molecular" graph G.

2 Graph theory

A simple, connected and undirected graph $G = \big(V(G),\ E(G)\big)$ is composed of a finite nonempty vertex set $V(G) = \{v_1,\ \ldots,\ v_p\}$ of $p \in \mathbb{N}$ vertices v_i and a finite edge set $E(G) = \{e_1,\ \ldots,\ e_q\}$ comprising $q \in \mathbb{N}$ distinct unordered pairs $e_\alpha = \{v_i,\ v_j\}$ of distinct vertices such that each element from $V(G)$ appears at least once [13]. The number of vertices $p = |V(G)|$ and edges $q = |E(G)|$ are called the order and size of G, respectively. By restricting edges to ensure bond uniqueness (any vertex pair is joined by at most one edge) and forbid loops (each edge must join distinct vertices), molecular structures are naturally represented by such graphs G. Each vertex $v_i \in V(G)$ has a nonempty neighbor set $S_i(G) = \{v_j : \{v_i,\ v_j\} \in E(G)\}$ that lists its adjacent vertices $v_j \in V(G)$, and $\deg(v_i) = |S_i(G)|$ is called the degree of v_i. The $p \times p$ square, symmetric and binary vertex adjacency matrix $\mathbf{A}(G)$ of a graph G, defined by

$$A_{ij}(G) = \begin{cases} 1, & \text{if } \{v_i,\ v_j\} \in E(G) \\ 0, & \text{otherwise} \end{cases},$$

is in one-to-one correspondence with the molecular structure and is arguably the most natural algebraic representation of topological connectivity. An elementary inductive argument [14] establishes that $\big(A^m\big)_{ij}(G)$ is the total number of m length walks (a sequence of vertices joined consecutively by edges) between vertices v_i and v_j. In particular, the degree of vertex v_i corresponding to the valency of atom i in the molecular structure is

$$\deg(v_i) = \big(A^2\big)_{ii}(G) = \sum_{j=1}^{p} \big(A_{ij}(G)\big)^2 = \sum_{j=1}^{p} A_{ij}(G) = \sum_{\alpha=1}^{q} Z_{i\alpha}(G),$$

so that the $p \times p$ diagonal degree matrix becomes

$$\mathbf{D}(G) = \sum_{i=1}^{p} \deg(v_i)\mathbf{e}_i\ \mathbf{e}_i^{\mathrm{T}} = \mathrm{diag}\Big(\deg(v_1),\ \ldots,\ \deg(v_p)\Big),$$

where the $\mathbf{e}_i$ are natural basis vectors in the coordinate vector space $\mathbb{R}^p$. Another useful representation of graph connectivity is the $p \times q$ binary vertex-edge incidence matrix $\mathbf{Z}(G)$ defined by

$$Z_{i\alpha}(G) = \begin{cases} 1, & \text{if } v_i \in e_\alpha \\ 0, & \text{otherwise} \end{cases}.$$

For each graph G there is an associated line graph $L(G)$ (also called the "derivative" graph [15]) such that $V\big(L(G)\big)$ is in bijective correspondence with $E(G)$ and

$$\mathbf{A}\big(L(G)\big) = \mathbf{Z}^{\mathrm{T}}(G)\,\mathbf{Z}(G) - 2\mathbf{I}_q, \tag{2}$$

where $\mathbf{I}_q$ is the $q \times q$ identity matrix. In other words, each edge of G is mapped to a vertex of $L(G)$, while two vertices of the line graph are adjacent if and only if their corresponding edges are incident in G (that is, they share a common endpoint). Furthermore, the associated Kirchhoff matrix $\mathbf{K}(G) = \mathbf{D}(G) - \mathbf{A}(G)$ (also called the Laplacian of G) satisfies a similar relationship

$$-\mathbf{K}(G) = \mathbf{Z}(G)\,\mathbf{Z}^{\mathrm{T}}(G) - 2\mathbf{D}(G),$$

and the positive semidefinite *signless* Laplace matrix of G is [16]

$$\mathbf{Q}(G) = \mathbf{Z}(G)\,\mathbf{Z}^{\mathrm{T}}(G) = \mathbf{A}(G) + \mathbf{D}(G).$$

The spectrum of $\mathbf{K}(G)$ provides a useful consistency check since the algebraic multiplicity of the zero eigenvalue is equal to the number of connected components in G [16]. Hence, for a physically sensible molecular graph G, the rank of $\mathbf{K}(G)$ must be $p - 1$.

We recall the following definitions and terminology. A cycle graph C_r comprises $p = r$ vertices, all of degree 2, connected in a closed chain by $q = r$ edges. Removing a single edge produces a path graph P_r (of order $p = r$ and size $q = r - 1$) with two terminal vertices of degree 1. The complete graph K_r on $p = r$ vertices is maximally connected with $q = \frac{1}{2}r(r-1)$ edges such that $\{v_i,\ v_j\} \in E\big(K_r\big)$ for all distinct $v_i,\ v_j \in V\big(K_r\big)$. A graph $G = \big(V(G),\ E(G)\big)$ is k-partite if the vertices can be partitioned into k disjoint sets, so that $V(G) = \cup_{r=1}^{k} V_r(G)$ where $V_r(G) \cap V_s(G) = \emptyset$ for $r \neq s$. The complete bipartite graph ($k = 2$) is denoted $K_{m,n}$ with $V\big(K_{m,n}\big) = V_1\big(K_{m,n}\big) \cup V_2\big(K_{m,n}\big)$ and size $p = m + n$ such that $m = \big|V_1\big(K_{m,n}\big)\big|$ and $n = \big|V_2\big(K_{m,n}\big)\big|$.

We will also have occasion to consider directed graphs $G = \big(V(G),\ A(G)\big)$ where edges are replaced by arrows specified by ordered pairs $\big(v_i,\ v_j\big) \in A(G)$ and oriented with the tail at vertex v_i pointing towards the head at vertex v_j. Associated with each vertex $v_i \in V(G)$ are two disjoint neighbor sets $S_i^-(G) = \big\{v_j : \big(v_i,\ v_j\big) \in A(G)\big\}$ and $S_i^+(G) = \big\{v_j\big(v_j,\ v_i\big) \in A(G)\big\}$ where $S_i = S_i^- \cup S_i^+$. At most one of S_i^- or S_i^+ may be empty. The indegree of vertex $v_i \in V(G)$ is $\deg^-\big(v_i\big) = \big|S_i^-(G)\big|$ and the outdegree is $\deg^+\big(v_i\big) = \big|S_i^+(G)\big|$. If $S_i^- = \emptyset$ so that $\deg^-\big(v_i\big) = 0$ then vertex v_i is

called a source and, similarly, if $S_i^+ = \emptyset$ so that $\deg^+(v_i) = 0$ then vertex v_i is called a sink.

3 Intramolecular interactions

In a molecular structure, covalently bonded atom pairs are considered to be adjacent. Similarly, a pair of adjacent bonds sharing a common "hinge" atom form a more or less flexible bend. The corresponding 3-body interaction is often described by an effective potential in terms of the external bond angle θ supplementary to the angle subtended at the hinge atom [17]. A 4-body dihedral interaction is associated with a pair of adjacent bends that share a common bond. Two situations are possible [18] (see Fig. 1): "proper" torsions arise when both hinge atoms are distinct, so that one bend is rotated about the other through a dihedral angle ϕ; while "improper" dihedral interactions link two bends through a common hinge atom, and are defined by a wag angle ω. Proper torsions typically account for geometric restrictions conferred by implicit substituents (usually protons) or lone electron pairs and may be alternatively characterized by a bond lying along the dihedral axis. Conversely, the dihedral axis of an improper torsion does not contain a bond and these interactions are used to constrain planar groups (like rings) or to hinder interconversion of stereocenters.

In graph theoretical terms, this hierarchical organization of interactions is precisely captured by iterated application of the line graph construction inductively defined by

$$L^n(G) = \begin{cases} G\,, & \text{if} \quad n = 0 \\ L\left(L^{n-1}(G)\right), & \text{if} \quad n > 0 \end{cases}. \tag{3}$$

A graph G establishes adjacency of vertices (i.e., bonds); the graph $L(G)$ encodes the adjacency of bonds (i.e., bends) in G; the graph $L\left(L(G)\right) = L^2(G)$ encodes the adjacency of bends (i.e., dihedrals) in G; and so on. Generally, the graph $L^{\nu-2}(G)$ encodes the adjacency of ν-body interactions in G. It has been shown [19] that the sequence of graphs $L^n(G)$ with $n = 0,\ 1,\ 2,\ \ldots$ has only four possible outcomes as $n \to \infty$:

1. if $G \cong C_r$ (a cycle graph on r vertices), then $L^n(G) \cong G$ for all $n \in \mathbb{N}$ (cycle graphs are the only connected graphs for which $L(G)$ is isomorphic to G);
2. if $G \cong K_{1,3}$ (the complete bipartite "claw" graph), then $L^n(G) \cong C_3$ (a triangle) for all $n \in \mathbb{N}$;
3. if $G \cong P_r$ (a path graph on r vertices), then $L^n(G) \cong P_{\max\{0,r-n\}}$ so each subsequent graph is a shorter path until eventually the sequence terminates at the trivial null graph;
4. otherwise, G is a "prolific" graph [20] so that the sizes of the graphs in the sequence eventually increase without bound,

$$\left|V\left(L^n(G)\right)\right| \to \infty \quad \text{as} \quad n \to \infty.$$

Ghebleh and Khatirinejad [21] have proved an interesting and chemically relevant result concerning the smallest non-negative integer m such that $L^m(G)$ is nonplanar:

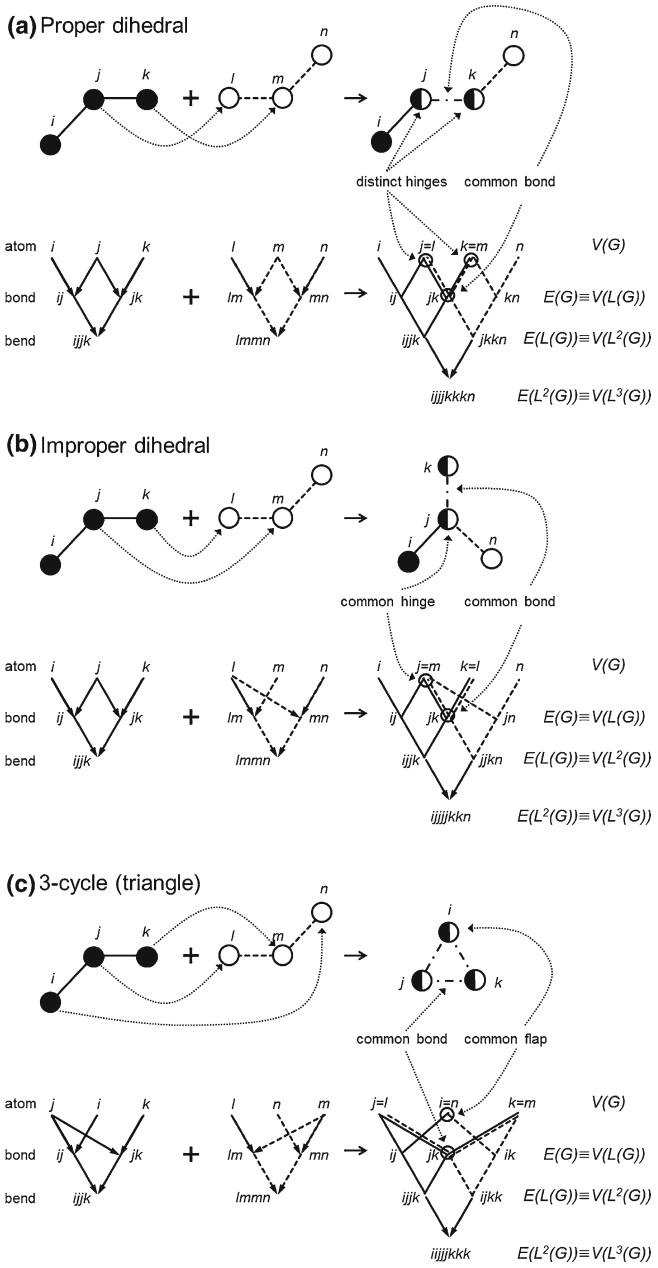

Fig. 1 Generic structures of 4-body interactions represented on the directed acyclic graph $H_v(G)$ are compared with the physical formation from the combination of two bends with atomic indexes ijk and lmn respectively. **a** Construction of the typical proper dihedral index sequence $ijjjkkkn$ comprising two triplet repeats arising from the distinct hinge atoms j and k that form the shared bond. A similar sequence $iijjjkkk$ arises for the degenerate 3-cycle structure as shown in (**c**), but where the "flap" atoms are also identified to close the ring. **b** The typical improper dihedral index sequence $ijjjjkkn$ with only a single hinge atom j

the so-called line index $m = \xi(G)$ of a graph G. In particular, if G is not prolific it is easy to see that $L^n(G)$ is planar for all $n \geq 0$, but for G prolific then $0 \leq \xi(G) \leq 4$ and a complete characterization of these graphs is possible [21].

3.1 Enumeration

Given a molecular graph G, the total number of intramolecular interactions $N_n(G)$ involving $n \in \mathbb{N}$ connected atoms (vertices on G) is generally given by

$$N_n(G) = \left|V\left(L^{n-1}(G)\right)\right| = \begin{cases} \left|V(G)\right|, & (n = 1) \\ \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{A}^2\left(L^{n-2}(G)\right)\right), & (n = 2,\ 3,\ 4,\ \ldots). \end{cases}$$

Here, $N_1(G)$ simply counts the number of 1-body interactions in the presence of an external field. Elementary combinatorial arguments also establish the handshaking lemma [22]: the number of bonds is just half the total number of incident edges over all vertices so that

$$N_{\mathrm{bond}}(G) = \frac{1}{2}\sum_{v\in V(G)} \deg(v) = \tfrac{1}{2}\,\mathrm{Tr}\left(\mathbf{D}(G)\right) = \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}(G)\mathbf{u} = N_2(G),$$

where $\mathbf{u} = \sum_{i=1}^{p} \mathbf{e}_i$ is a p-vector of ones. By summing the number of possible bond pairs over each vertex, the total bend count is obtained

$$\begin{aligned} N_{\mathrm{bend}}(G) &= \sum_{v\in V(G)} \binom{\deg(v)}{2} = \frac{1}{2}\sum_{v\in V(G)} \deg(v)\big(\deg(v)-1\big) \\ &= \tfrac{1}{2}\,\mathrm{Tr}\left(\mathbf{D}(G)\big(\mathbf{D}(G)-\mathbf{I}\big)\right) = \tfrac{1}{2}\,\mathrm{Tr}\left(\mathbf{D}^2(G)\right) - N_2(G) \\ &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}^2(G)\mathbf{u} - N_2(G) = N_3(G). \end{aligned}$$

Similarly reckoning the combinations of bond triplets over all vertices yields the number of improper dihedral interactions

$$\begin{aligned} N_{\mathrm{impr}}(G) &= \sum_{v\in V(G)} \binom{\deg(v)}{3} = \frac{1}{3!}\sum_{v\in V(G)} \deg(v)\big(\deg(v)-1\big)\big(\deg(v)-2\big) \\ &= \tfrac{1}{6}\mathbf{u}^{\mathrm{T}}\mathbf{D}^3(G)\mathbf{u} - N_3(G) - \tfrac{1}{3}N_2(G). \end{aligned}$$

Other possible arrangements of three contiguous bonds are just the proper dihedrals and triangular 3-cycles. In total, these arrangements can be enumerated as follows: for each bond, calculate the product of the number of free edges otherwise incident

on each of the two vertices; then sum over all edges to get

$$\begin{aligned} N_{\text{prop}}(G) + 3N_{3\text{cyc}}(G) &= \sum_{\{v_i,\ v_j\}\in E(G)} \Big(\deg\big(v_i\big) - 1\Big)\Big(\deg\big(v_j\big) - 1\Big) \\ &= \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\big(D_{ii}(G) - 1\big)A_{ij}(G)\big(D_{jj}(G) - 1\big) \\ &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}(G)\mathbf{A}(G)\mathbf{D}(G)\mathbf{u} - 2N_3(G) - N_2(G), \end{aligned}$$

where we have observed the identity $\sum_{j=1}^{p} A_{ij}(G) = D_{ii}(G)$. The number of 3-cycles, however, can be obtained directly from the adjacency matrix as

$$N_{3\text{cyc}}(G) = \tfrac{1}{3!}\,\mathrm{Tr}\,\big(\mathbf{A}^3(G)\big).$$

The total 4-body interaction number N_4 includes torsions of both types as well as any 3-cycles present

$$N_4(G) = N_{\text{prop}}(G) + 3N_{\text{impr}}(G) + 3N_{3\text{cyc}}(G),$$

but over-counts the improper dihedrals and triangles by distinguishing the three rotational permutations of labels.

Combining these results gives the useful consistency checksums

$$\begin{aligned} N_1(G) &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{A}(G)\mathbf{u}, \\ N_2(G) &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}(G)\mathbf{u}, \\ N_3(G) &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}^2(G)\mathbf{u} - N_2(G), \\ N_4(G) &= \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}(G)\mathbf{Q}(G)\mathbf{D}(G)\mathbf{u} - 5N_3(G) - 2N_2(G), \end{aligned}$$

that involve only the molecular graph G.

3.2 Indexing

By definition of the line graph transformation, for $A_{ij}\big(L^n(G)\big) = 1$, the indexes $i = \alpha$ and $j = \beta$ point to the edges $e_\alpha,\ e_\beta \in E\big(L^{n-1}(G)\big)$ that share a common vertex $v_l \in V\big(L^{n-1}(G)\big)$. In turn, these edges $e_\alpha = \big\{v_k,\ v_l\big\}$ and $e_\beta = \big\{v_l,\ v_m\big\}$ are associated with the adjacency matrix of the previous graph in the recursive hierarchy so that $A_{kl}\big(L^{n-1}(G)\big) = A_{lm}\big(L^{n-1}(G)\big) = 1$. Successively backtracking to the source graph G yields, for each order n, a sequence of atomic indexes that participate in an n-body interaction. A formal pseudo-code implementation of this procedure is set out in Algorithm 1. Details of each sequence structure completely characterizes the interaction form.

For concreteness, specify a maximum order ν for the multibody interactions of interest so that n-body terms with $n = 1,\ 2,\ \ldots,\ \nu$ are considered. Most commonly in

Algorithm 1: Generate ν-body atom index list.

Function `List(`ν, p, $\mathbf{A}$`)`

Data: $\nu \in \mathbb{N}$ the order of multibody interactions.

$p = |V(G)| \in \mathbb{N}$, the order of simply connected graph $G = (V,\ E)$.

$\mathbf{A}(G)$ a $p \times p$ symmetric binary matrix.

Result: $r = \left|V\left(L^{\nu-1}(G)\right)\right| \in \mathbb{N}$, the order of the $\nu - 1$ iterated line graph $L^{\nu-1}(G)$ of simply connected graph $G = (V,\ E)$.

edges, an r-length vector listing (for $\nu > 1$) the edge set $E\left(L^{\nu-2}(G)\right)$ with elements expressed as a nested hierarchy of order pairs identifying the vertices of simply connected graph $G = (V,\ E)$ associated with a ν-body intramolecular interaction. If $\nu = 1$, then edges $= V(G)$.

$\mathbf{A}_{\mathrm{L}}\left(L^{\nu-1}(G)\right)$ a $r \times r$ symmetric binary matrix.

```
if ν > 1 then
    (q, e_old, M) ← List(ν − 1, p, A)                    /* recursive call */
    α ← 0                                                /* initialize edge count */
    for j = 1 to q do                                    /* loop over columns */
        for i = 1 to j − 1 do                            /* loop over rows */
            if M[i, j] = 1 then                          /* edge e_α = {v_i, v_j} here */
                α ← α + 1                                /* increment edge count */
                edges[α] ← (e_old[i], e_old[j])          /* construct ordered pair */
    r ← α                                                /* store edge count */
    Z[q, r] ← 0                                          /* initialize incidence matrix */
    α ← 0                                                /* initialize edge count */
    for j = 1 to q do                                    /* loop over columns */
        for i = 1 to j − 1 do                            /* loop over rows */
            if M[i, j] = 1 then                          /* edge e_α = {v_i, v_j} here */
                α ← α + 1                                /* increment edge counter */
                Z[i, α] ← 1                              /* set e_α */
                Z[j, α] ← 1
    A_L ← Z^T Z − 2I_r                                   /* equation 2 */
else                                                     /* base case */
    for i = 1 to p do                                    /* loop over rows */
        if A[i, i] ≠ 0 then break                        /* EXCEPTION: G not simple (loop) */
        edges[i] ← i                                     /* vertex of G */
    r ← p
    A_L ← A
return (r, edges, A_L)
```

molecular applications $\nu = 4$. A convenient representation of the recursive structure defined by (3) is itself a weakly connected ν-partite directed acyclic graph (DAG) denoted $H_\nu(G)$ on the union of disjoint vertex sets $\cup_{n=1}^{\nu} V_n(G)$ where $V_n(G) = V\left(L^{n-1}(G)\right)$. For clarity, define the vertex set $V\left(H_\nu(G)\right)$ of $H_\nu(G)$ with elements $v_{jn} \in V_n(G)$. The edge directions induce a partial order relation on the vertices so that $v_{im} \preceq v_{jn}$ only if $m < n$ where $v_{im} \in V_m(G)$ and $v_{jn} \in V_n(G)$. Sources of $H_\nu(G)$ with indegree 0 are just the atomic indexes of the molecular graph G. All other vertices on $H_\nu(G)$ have indegree 2 as a consequence of the defining property of an edge in $L^n(G)$. Similarly, the sinks of $H_\nu(G)$ with outdegree 0 are just the vertices

of $L^{\nu-1}(G)$. At each level $n = 1, 2, \ldots, \nu - 1$ all other vertices $v_{jn} \in V_n(G)$ on $H_\nu(G)$ have outdegrees given by $\mathbf{D}\big(L^{n-1}(G)\big)$. Moreover, the hierarchy of line graphs provides a natural topological order on $H_\nu(G)$. Further, the adjacency matrix takes the block tridiagonal form

$$\mathbf{A}\big(H_\nu(G)\big) = \begin{pmatrix} \mathbf{0} & \mathbf{Z}(G) & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{Z}^{\mathrm{T}}(G) & \mathbf{0} & \mathbf{Z}\big(L(G)\big) & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}^{\mathrm{T}}\big(L(G)\big) & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}\big(L^{\nu-2}(G)\big) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}^{\mathrm{T}}\big(L^{\nu-2}(G)\big) & \mathbf{0} \end{pmatrix}.$$

By virtue of this DAG structure, each vertex $v_{jn} \in V\big(H_\nu(G)\big)$ is associated with a unique sequence of 2^{n-1} atomic indexes v_{i1} that define each n-body interaction. For the trivial cases $n = 1$ and $n = 2$, these sequences correspond respectively with the individual atom list $V(G)$ and the atom pairs $E(G)$ defining the molecular bonds. Bends ($n = 3$) are signified by a pattern of $2^2 = 4$ atoms with a single repeated index identifying the common hinge atom. An 8-atom sequence classifies a 4-body interaction and distinguishes between three types as illustrated in Fig. 1. A pair of triply repeated atom indexes identifies the distinct hinge atoms of a proper torsion, provided the remaining two indexes are different (Fig. 1a). Otherwise, a common pair of "flap" atoms signals a degenerate 3-cycle where only three atom indexes appear (Fig. 1c). Each triangle generates ${}^3C_2 = 3$ such sequences by the arbitrary choice of a single common bond. The single common hinge of an improper dihedral corresponds to a fourfold repeated index among four distinct labels (Fig. 1b). Again, three sequences are generated for each improper dihedral by arbitrary assignment of the shared bond.

4 Examples

4.1 A toy model: methylcyclopropane

The molecular graph G indicated in Fig. 2 presents amongst other possibilities a plausible lumped model for methylcyclopropane where hydrogen atoms are absorbed onto the carbon backbone in the usual way. Direct inspection of G immediately establishes four 1-body interactions in the presence of an external field (that is, the number of "atoms" $N_1 = 4$) and also four 2-body bonds $\big(N_2 = 4\big)$. Atom 2 is the hinge for three distinct 3-body bends with a further two hinged at atoms 3 and 4 respectively, to give $N_3 = 5$. Among the 4-body interactions there are two proper torsions $\big(N_{\mathrm{prop}} = 2\big)$ with distinct hinge atoms 2-3 and 2-4, respectively, as well as a single improper dihedral $\big(N_{\mathrm{impr}} = 1\big)$ with common hinge atom 2. A single 3-cycle is present $\big(N_{3\mathrm{cyc}} = 1\big)$ so that $N_4 = 2 + 3 \times (1 + 1) = 8$. Adjacency matrices for the iterated line graphs necessary for describing interactions up to the 4-body level are given by

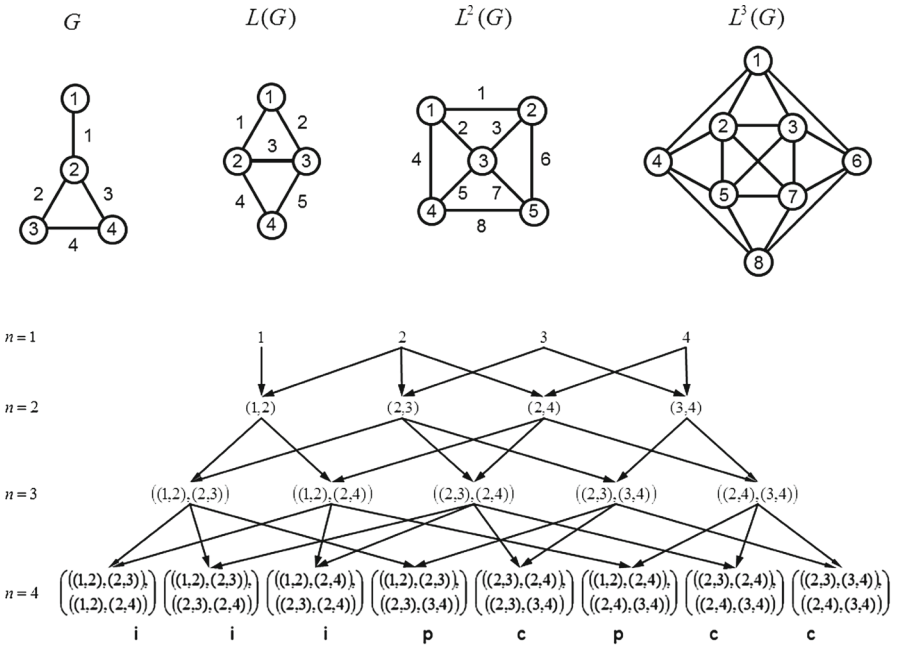

Fig. 2 A plausible effective model of methylcyclopropane is represented by the graph G. The iterated line graphs $L^{n-1}(G)$ are also shown for $n = 2,\ 3,\ 4$. Vertices $v_{im} \in V_m(G) = V\left(L^{m-1}(G)\right)$ and edges $e_{km} = \left\{v_{im},\ v_{jm}\right\}$ are labelled. The corresponding directed acyclic graph (DAG) generated by the line graph hierarchy is given where the vertices at each level n are denoted by the inherited sequence of atomic labels from G as indicated by the directed edges. Inspection of G confirms the 4-body interactions identified by the DAG sinks and comprise of: two proper torsions (denoted "p") with distinct hinge atoms 2-3 and 2-4, respectively; a single improper dihedral (denoted "i") with common hinge 2; and a single 3-cycle (denoted "c") of atoms 2, 3 and 4

$$\mathbf{A}(G) = \begin{pmatrix} 0 & 1^1 & 0 & 0 \\ 1^1 & 0 & 1^2 & 1^3 \\ 0 & 1^2 & 0 & 1^4 \\ 0 & 1^3 & 1^4 & 0 \end{pmatrix}, \quad \mathbf{A}\left(L(G)\right) = \begin{pmatrix} 0 & 1^1 & 1^2 & 0 \\ 1^1 & 0 & 1^3 & 1^4 \\ 1^2 & 1^3 & 0 & 1^5 \\ 0 & 1^4 & 1^5 & 0 \end{pmatrix},$$

$$\mathbf{A}\left(L^2(G)\right) = \begin{pmatrix} 0 & 1^1 & 1^2 & 1^4 & 0 \\ 1^1 & 0 & 1^3 & 0 & 1^6 \\ 1^2 & 1^3 & 0 & 1^5 & 1^7 \\ 1^4 & 0 & 1^5 & 0 & 1^8 \\ 0 & 1^6 & 1^7 & 1^8 & 0 \end{pmatrix}.$$

For added clarity, the edge index associated with each adjacent vertex pair is indicated by the superscript. From these matrices, the DAG obtained that represents the line graph hierarchy is shown in Fig. 2. The atomic index sequences automatically generated on the DAG, particularly at the 4-body level ($n = 4$), confirm the informal analysis of the molecular graph. It is easy to show that the complete graph on five vertices K_5 is a minor of $L^3(G)$, whence it follows from the theorem of Wagner [23] that $L^3(G)$ is

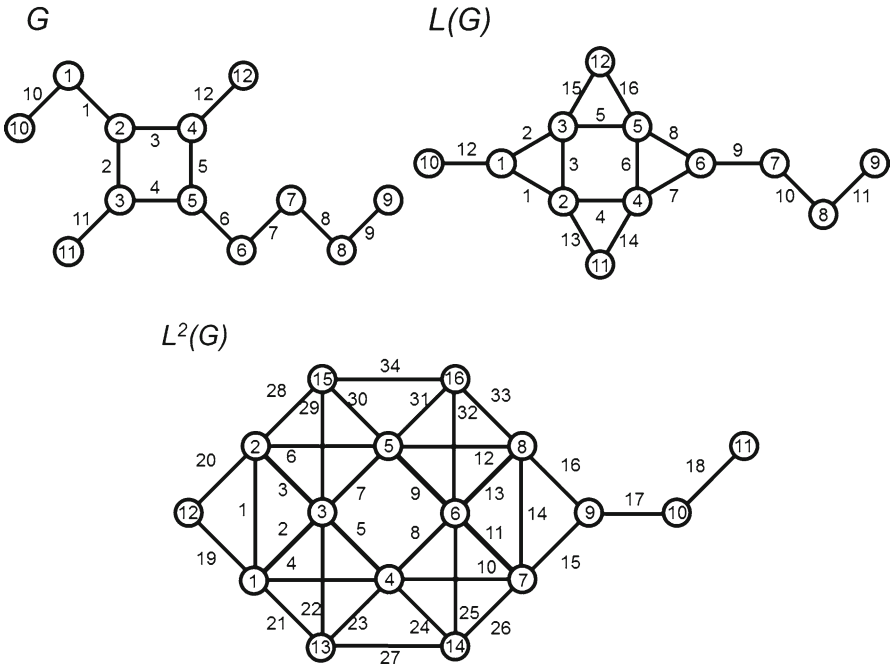

Fig. 3 Effective coarse grained bile salt model of Vila Verde and Frenkel [24]. The molecular graph G is shown along with the iterated line graphs $L(G)$ and $L^2(G)$. Vertices and edges are labelled

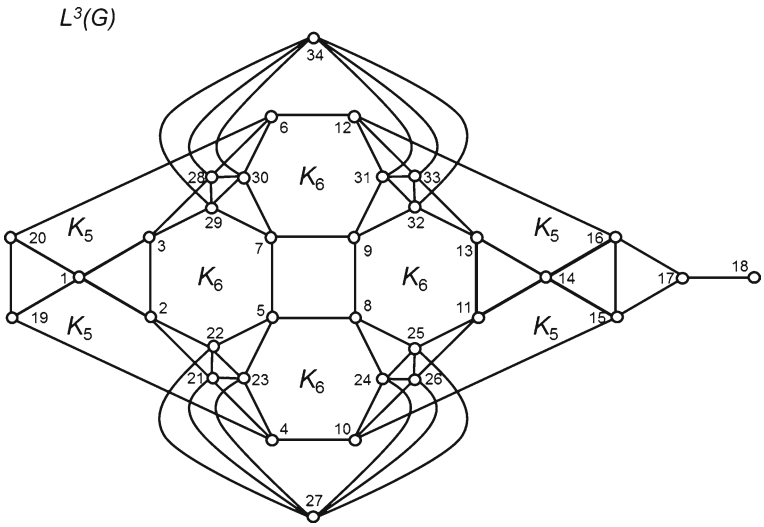

Fig. 4 Effective coarse grained bile salt model of Vila Verde and Frenkel [24]. The molecular line graph $L^3(G)$. To simplify the diagram, subgraphs corresponding to complete graphs K_q on q vertices (q-cliques of $L^3(G)$) have been rendered as pentagons ($q = 5$) and hexagons ($q = 6$)

Table 1 Vertex sequences associated with the directed acyclic graph (DAG) description of the line graph hierarchy for the effective bile salt model of Vila Verde and Frenkel [24]

N_n	$n = 1$	$n = 2$	$n = 3$	$n = 4$	
1	1	(1, 2)	((1, 2), (2, 3))	(((1, 2), (2, 3)), ((1, 2), (2, 4)))	i
2	2	(2, 3)	((1, 2), (2, 4))	(((1, 2), (2, 3)), ((2, 3), (2, 4)))	i
3	3	(2, 4)	((2, 3), (2, 4))	(((1, 2), (2, 4)), ((2, 3), (2, 4)))	i
4	4	(3, 5)	((2, 3), (3, 5))	(((1, 2), (2, 3)), ((2, 3), (3, 5)))	p
5	5	(4, 5)	((2, 4), (4, 5))	(((2, 3), (2, 4)), ((2, 3), (3, 5)))	p
6	6	(5, 6)	((3, 5), (4, 5))	(((1, 2), (2, 4)), ((2, 4), (4, 5)))	p
7	7	(6, 7)	((3, 5), (5, 6))	(((2, 3), (2, 4)), ((2, 4), (4, 5)))	p
8	8	(7, 8)	((4, 5), (5, 6))	(((2, 3), (3, 5)), ((3, 5), (4, 5)))	p
9	9	(8, 9)	((5, 6), (6, 7))	(((2, 4), (4, 5)), ((3, 5), (4, 5)))	p
10	10	(1, 10)	((6, 7), (7, 8))	(((2, 3), (3, 5)), ((3, 5), (5, 6)))	p
11	11	(3, 11)	((7, 8), (8, 9))	(((3, 5), (4, 5)), ((3, 5), (5, 6)))	i
12	12	(4, 12)	((1, 2), (1, 10))	(((2, 4), (4, 5)), ((4, 5), (5, 6)))	p
13			((2, 3), (3, 11))	(((3, 5), (4, 5)), ((4, 5), (5, 6)))	i
14			((3, 5), (3, 11))	(((3, 5), (5, 6)), ((4, 5), (5, 6)))	i
15			((2, 4), (4, 12))	(((3, 5), (5, 6)), ((5, 6), (6, 7)))	p
16			((4, 5), (4, 12))	(((4, 5), (5, 6)), ((5, 6), (6, 7)))	p
17				(((5, 6), (6, 7)), ((6, 7), (7, 8)))	p
18				(((6, 7), (7, 8)), ((7, 8), (8, 9)))	p
19				(((1, 2), (2, 3)), ((1, 2), (1, 10)))	p
20				(((1, 2), (2, 4)), ((1, 2), (1, 10)))	p
21				(((1, 2), (2, 3)), ((2, 3), (3, 11)))	p
22				(((2, 3), (2, 4)), ((2, 3), (3, 11)))	p
23				(((2, 3), (3, 5)), ((2, 3), (3, 11)))	i
24				(((2, 3), (3, 5)), ((3, 5), (3, 11)))	i
25				(((3, 5), (4, 5)), ((3, 5), (3, 11)))	p
26				(((3, 5), (5, 6)), ((3, 5), (3, 11)))	p
27				(((2, 3), (3, 11)), ((3, 5), (3, 11)))	i
28				(((1, 2), (2, 4)), ((2, 4), (4, 12)))	p
29				(((2, 3), (2, 4)), ((2, 4), (4, 12)))	p
30				(((2, 4), (4, 5)), ((2, 4), (4, 12)))	i
31				(((2, 4), (4, 5)), ((4, 5), (4, 12)))	i
32				(((3, 5), (4, 5)), ((4, 5), (4, 12)))	p
33				(((4, 5), (5, 6)), ((4, 5), (4, 12)))	p
34				(((2, 4), (4, 12)), ((4, 5), (4, 12)))	i

At each level n, the inherited sequence of atomic labels from the molecular graph G is listed. Inspection of G confirms the 4-body interactions identified by the DAG sinks and comprise of: 22 proper torsions (denoted "p") and 4 improper dihedrals (denoted "i") with no 3-cycles

nonplanar. All of G, $L(G)$ and $L^2(G)$ are manifestly planar (see Fig. 2) so the line index $\xi(G) = 3$ in accord with the result of Ghebleh and Khatirinejad [21].

4.2 Bile salt: taurocholate

A chemically relevant example, central to the hydrolysis and solubilisation of lipid associated with food digestion in the human lower gastrointestinal tract, are the bile salt and bile acid surfactants. Vila Verde and Frenkel [24] have proposed an effective coarse-grained model of trihydroxy bile salts (taurocholate) for a molecular dynamics study of micelle formation that is related to the rate and extent of nutrient absorption by intestinal cells. The authors proposed a "three-to-one" mapping scheme, that groups three carbon or nitrogen atoms into a single bead, to arrive at the molecular graph G shown in Fig. 3. Iterated line graphs up to $L^3(G)$ are collected in Figs. 3 and 4. Table 1 lists the vertices for the corresponding DAG description of the hierarchy as sequences of bead indexes.

Overall, the counts of bonds, bends, proper torsions and improper dihedrals are obtained as follows,

$$N_{\text{bond}} = 12, \quad N_{\text{bend}} = 16, \quad N_{\text{prop}} = 22, \quad N_{\text{impr}} = 4,$$

where $N_4 = 22 + 3 \times 4 = 34$. Clearly, there are no 3-cycles in this example so $N_{\text{cyc}} = 0$. It is easy to verify that $L^2(G)$ admits the minor K_5 and hence, by Wagner's theorem [23], it follows that $L^2(G)$ is nonplanar. Both G and $L(G)$ are manifestly planar so the line index $\xi(G) = 2$ in accord with the result of Ghebleh and Khatirinejad [21].

5 Conclusion

The line graph transformation provides a practical and elegant theoretical tool for exhaustively enumerating and indexing many-body intramolecular interactions. Given a suitable graphical representation of a molecular structure, an explicit pseudo-code implementation of the recursive line graph algorithm is given for automatically generating complete canonical lists of atomic indexes associated with each interaction order. No attempt has been made to computationally optimize this algorithm or the associated data structures. Instead, clarity of exposition is the main objective here. We anticipate the main application will involve embedding the algorithm within a Monte Carlo or Molecular Dynamics simulation code where other implementation details will determine the most efficient realization. In accord with common practice, intramolecular interactions up to order 4 have been considered (bonds, bends and dihedrals), but the method can be extended to arbitrarily many atomic centers. Higher order interactions will involve increasingly many sub-type variations and polycyclic structures. Two specific examples are discussed: a toy model of methylcyclopropane and a published effective potential model of taurocholate bile salt [24] that is relevant for the study of digestive processes in the human lower gastrointestinal tract.

Acknowledgments This work was financially supported by the Biotechnology and Biological Sciences Research Council through its core strategic grant to the Institute of Food Research. We also thank Andrew Watson for critical reading of the manuscript.

References

1. D. Bonchev, J. Mol. Struct. (Theochem) **336**, 137 (1995)
2. D.H. Rouvray, A.T. Balaban, in *Applications of Graph Theory*, ed. by R.J. Wilson, L.W. Beinecke (Academic Press, London, 1979), p. 177
3. H.N.V. Temperley, in *Applications of Graph Theory*, ed. by R.J. Wilson, L.W. Beinecke (Academic Press, London, 1979), p. 121
4. J.-P. Hanson, I.R. McDonald, *Theory of Simple Liquids*, 2nd edn. (Academic Press, London, 1986), pp. 79–92
5. B. Kirchner, Phys. Rep. **440**, 1 (2007)
6. J. Hutter, WIRE: Comp. Mol. Sci. **2**, 604 (2012)
7. R.D. Kohn, A.D. Becke, R.G. Parr, J. Phys. Chem. **100**, 12974 (1996)
8. R. Penfold, S. Abbas, S. Nordholm, Fluid Phase Equilib. **120**, 39 (1996)
9. J.T. Padding, A.A. Louis, Phys. Rev. E **74**, 031402 (2006)
10. M.P. Allen, in *Computational Soft Matter: From Synthetic Polymers to Proteins*, ed. by N. Attig, K. Binder, H. Grubmüller, K. Kremer (John von Neumann Institute for Computing, Jülich, 2004), p. 1
11. S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. de Vries, J. Phys. Chem. B. **111**, 7812 (2007)
12. R.L. Hemminger, L.W. Beinecke, in *Selected Topics in Graph Theory*, ed. by L.W. Beinecke, R.J. Wilson (Academic Press, London, 1978), p. 271
13. F. Harary, *Graph Theory* (Addison-Wesley, Reading, 1969)
14. J. Clark, D.A. Holton, *First Look at Graph Theory* (World Scientific, Singapore, 1991)
15. G. Sabidussi, Math. Zeitschr **76**, 385 (1961)
16. A.E. Brouwer, W.H. Haemers, *Spectra of Graphs* (Springer, New York, 2012)
17. P.J. Flory, *Statistical Mechanics of Chain Molecules* (Carl Hanser Verlag, Munich, 1989)
18. R.E. Tuzun, D.W. Noid, B.G. Sumpter, J. Comp. Chem. **18**, 1513 (1997)
19. A.C.M. van Rooij, H.S. Wilf, Acta Math. Hungar. **16**, 263 (1965)
20. M. Knor, P. Potočnik, R. Škrekovski, Discrete Appl. Math. **160**, 2234 (2012)
21. M. Ghebleh, M. Khatirinejad, Discrete Math. **308**, 144 (2012)
22. R.J. Wilson, *Introduction to Graph Theory*, 3rd edn. (Longman, Harlow, 1985)
23. R. Diestel, *Graph Theory*, 3rd edn. (Springer, Heidelberg, 2005)
24. A. Vila Verde, D. Frenkel, Soft Matter **6**, 3815 (2010)