



Institut für Numerische Simulation

Rheinische Friedrich-Wilhelms-Universität Bonn

Friedrich-Hirzebruch-Allee 7 • 53115 Bonn • Germany  
phone +49 228 73-69828 • fax +49 228 73-69847  
[www.ins.uni-bonn.de](http://www.ins.uni-bonn.de)

J. Barker, M. Griebel, J. Hamaekers

## On Multilevel Energy-Based Fragmentation Methods

INS Preprint No. 2403

November 2024



# On Multilevel Energy-Based Fragmentation Methods

James Barker<sup>a</sup>, Michael Griebel<sup>a,b</sup>, and Jan Hamaekers<sup>b,\*</sup>

<sup>a</sup>Institute for Numerical Simulation, University of Bonn, Friedrich-Hirzebruch-Allee 7, 53115 Bonn, Germany

<sup>b</sup>Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven 1,  
53757 Sankt Augustin, Germany

\*Corresponding author: [jan.hamaekers@scai.fraunhofer.de](mailto:jan.hamaekers@scai.fraunhofer.de)

March 28, 2025

Energy-based fragmentation methods approximate the potential energy of a molecular system as a sum of contribution terms built from the energies of particular subsystems. Some such methods reduce to truncations of the many-body expansion (MBE); others combine subsystem energies in a manner inspired by the principle of inclusion/exclusion (PIE). The combinatorial technique of Möbius inversion of sums over partially ordered sets, which generalizes the PIE, is known to provide a non-recursive expression for the MBE contribution terms, and has also been connected to related cluster expansion methods. We build from these ideas a very general framework for decomposing potential functions into energetic contribution terms associated with elements of particular partially ordered sets (posets) and direct products thereof. Specific choices immediately reproduce not only the MBE, but also a number of other existing decomposition forms, including, e.g., the multilevel ML-BOSSANOVA schema. Furthermore, a different choice of poset product leads to a setup familiar from the combination technique for high-dimensional approximation, which has a known connection to quantum-chemical composite methods. We present the ML-SUPANOVA decomposition form, which allows the further refinement of the terms of an MBE-like expansion of the Born-Oppenheimer potential according to systematic hierarchies of *ab initio* methods and of basis sets. We outline an adaptive algorithm for the *a posteriori* construction of quasi-optimal truncations of this decomposition. Some initial experiments are reported and discussed.

## 1. Introduction

The generally high-dimensional nature of the electronic Schrödinger equation renders it inaccessible to conventional numerical techniques. Some recourse is possible to the specialized mechanisms of computational quantum chemistry [1–5], but the involved costs still scale superlinearly, so these cannot be applied to larger molecular systems, especially not to an acceptable level of accuracy. A common tactic is to decompose the single overarching electronic problem into subproblems. Solutions to these can be approximated independently, and combined back into an approximate solution for the original problem. Distinct versions of this tactic are recognisably present in *composite methods* [6–8], and in *energy-based fragmentation methods* [7, 9–11].

The working formulae for fragmentation methods are often related to or inspired by the well-known *many-body expansion* (MBE) [11–14]. Extensions apply counting arguments founded in the principle of inclusion/exclusion (PIE) [13, 15]. Significant overlap exists here with those fragmentation methods

which construct and identify their fragments using ideas from graph theory [14, 16–22], and also with *multilevel fragmentation methods* [14, 20, 21, 23–27], which combine the results of subsystem calculations performed at differing levels of theory.

Much work has gone into understanding the properties of and relationships between different fragmentation methods, often with an eye to their unification [13, 15, 27–29]. We suggest that the combinatorial theory of *Möbius inversion* [30–32], which can be understood as an extension of inclusion/exclusion, provides a useful theoretical framework, not only for the comparison and analysis of existing fragmentation methods, but also for the development of new ones. The connection between Möbius inversion and the MBE is known [12, 33], although apparently unexplored in the modern fragmentation method context. But Möbius inversion has long been applied in cluster expansion methods, see, e.g., [33–35]. These are closely related to the MBE [33], and some are explicitly founded in graph theory, e.g., [34].

We consider here a very general decomposition form over direct products of partially ordered sets (*posets*). Specific choices of posets directly and deliberately reproduce existing decomposition forms well-known in mathematics. The simplest case would be a totally ordered set, i.e., a chain, and direct products of chains lead to grid-based decompositions like those underlying the *combination technique* [36–42]. The choice instead of a single Boolean algebra, that is, an inclusion-ordered powerset  $2^{[n]}$ , delivers an *ANOVA-like decomposition* [43, 44], cf. [45]. Any subposet of a Boolean algebra is isomorphic to an ordered subset of the induced subgraphs of some undirected graph  $G$  with vertex set  $[n]$  [46], and the structural properties of some such subsets make them either more or less appealing for our purposes. From the chemical perspective, the resulting decompositions either mimic or directly reconstruct, e.g., composite methods [6], the MBE, graph-based forms like the BOSSANOVA (*Bond-Order disSection ANOVA*) [18, 19] decomposition and particular cases of the CGTCE [34], and finally, multilevel mechanisms such as ML-BOSSANOVA [14] and ML-FCR [27]. A product of either a Boolean algebra or a subposet thereof with multiple chains extends these ideas to a new decomposition type that we call *ML-SUPANOVA*.

This article is based on [47]; further details including data, graphics, and experiments are given there. The structure is as follows: In Section 2, we progressively develop and motivate our new multilevel decomposition. Section 2.1 formalizes a hierarchy of model approximations to the Born-Oppenheimer potential function, much like those used in composite methods. Section 2.2 establishes some basics regarding fragmentation methods. Section 2.3 introduces the combinatorial tools that we need to build and operate our multilevel scheme, and demonstrates their application. In Section 2.4, we use these tools to uncover an issue in the BOSSANOVA setup, and consider the introduction of convex subgraphs as a partial panacea. Section 2.5 combines ideas from both composite methods and multilevel energy-based fragmentation methods, and introduces the final ML-SUPANOVA decomposition. We outline in Section 3 an adaptive *a posteriori* algorithm for generating cost-effective truncations of this decomposition. Some initial experimental results are discussed in Section 4, and we make some concluding remarks in Section 5.

## 2. Decomposition-Based Techniques in Computational Chemistry

### 2.1. A model hierarchy and composite methods

We begin by constructing for later use a formal hierarchy of familiar models in computational quantum chemistry. For general background, we refer to [1–5]. For a molecular system composed of  $M$  clamped nuclei and  $N$  electrons, the latter with spatial/spin coordinates  $\{x_i\}_{i=1}^N$ , we consider the usual electronic Schrödinger equation

$$H\Psi(x_1, \dots, x_N) = E\Psi(x_1, \dots, x_N). \quad (1)$$

We will refer to as the *Born-Oppenheimer potential* the function

$$V^{\text{BO}}(X_1, \dots, X_M) := \sum_{1 \leq A < B \leq M} \frac{Z_A Z_B}{\|R_B - R_A\|_2} + \inf_{\substack{\Psi \in \mathcal{V} \\ \langle \Psi, \Psi \rangle = 1}} \langle \Psi, H\Psi \rangle \quad (2)$$

delivering the ground-state total energy under the Born-Oppenheimer approximation, cf. [14]. This includes the repulsion energy contributed by the nuclei with coordinates  $\{X_A = (R_A \in \mathbb{R}^3, Z_A \in \mathbb{N})\}_{A=1}^M$ , where each  $R_A$  is a spatial coordinate and each  $Z_A$  is an atomic number. In (2),  $\mathcal{V}$  is an appropriate solution space for the weak form of (1), see [2, 48].



We restrict our attention to a collection of standard *ab initio* methods which build on the Hartree-Fock (HF) ansatz [1] in order to approximate the minimizing  $\Psi$  in (2). These involve a discretization of  $H^1(\mathbb{R}^3)$  via a practically finite *basis set* of  $N_{\text{AO}}$  atomic orbital functions, usually Gaussian-type [3–5, 49, 50]. The cost scaling behaviour of these methods depends on  $N_{\text{AO}}$ . For HF, the cost scales formally as  $\mathcal{O}(N_{\text{AO}}^4)$ , and for an additional estimate of the correlation energy from second-order perturbative Møller-Plesset theory (MP2) [1, 51], basically as  $\mathcal{O}(N_{\text{AO}}^5)$ . For better-quality estimates, we turn to coupled-cluster methods: CCSD, CCSDT, etc. [48, 52]. The cost of a general CC( $n$ ) coupled cluster calculation considering  $n$ th-order excitations [53] can be rather loosely given as  $\mathcal{O}(N_{\text{AO}}^{2n+2})$ , but see [53] for more precision. The perturbative approach underpinning the CCSD(T) [54] correction has been generalized to CC( $n$ )( $n+1$ ) correction terms applicable to CC( $n$ ) calculations [55–57]. Again loosely, these cost  $\mathcal{O}(N_{\text{AO}}^{2n+3})$  to evaluate, but see and cf. [8, 52, 56]. Finally, the best possible estimate of the correlation energy for a particular discretization of  $H^1(\mathbb{R}^3)$  is given by full configuration-interaction (FCI), at cost  $\sim \mathcal{O}(N_{\text{AO}}^N)$ , see, e.g., [48]. Taken all together, these methods present a hierarchy containing  $2N-1$  elements,<sup>1</sup> with regularly increasing cost and presumably increasing accuracy, cf., e.g., [8, 52]:

$$\begin{aligned} \text{HF}[\mathcal{O}(N_{\text{AO}}^4)] &\rightarrow \text{MP2}[\mathcal{O}(N_{\text{AO}}^5)] \rightarrow \text{CCSD}[\mathcal{O}(N_{\text{AO}}^6)] \rightarrow \text{CCSD(T)}[\mathcal{O}(N_{\text{AO}}^7)] \\ &\rightarrow \cdots \rightarrow \text{CC}(n)[\mathcal{O}(N_{\text{AO}}^{2n+2})] \rightarrow \text{CC}(n)(n+1)[\mathcal{O}(N_{\text{AO}}^{2n+3})] \\ &\rightarrow \cdots \rightarrow \text{FCI}[\sim \mathcal{O}(N_{\text{AO}}^N)]. \end{aligned} \quad (3)$$

In practice, the size  $N_{\text{AO}}$  of the discretizing basis set is adjustable only indirectly, as a consequence of the generative use of, e.g., one of the cc-pVnZ basis sets [58]. A different, but also regular quality/cost hierarchy is here exposed as  $n$  is varied. The error due to the incompleteness of the basis set is expected to decrease regularly as  $n \rightarrow \infty$  and the so-called *complete basis set* (CBS) limit is approached [50]. This regularity is supported at least in practice by the success of extrapolation procedures; see, e.g., [59], and discussion in [14]. In any case, the total number of basis functions required to describe an  $M$ -atom system with an cc-pVnZ basis set goes as  $N_{\text{AO}} \sim Mn^3$ ; see [4, Sec. 8.3] for precise counting formulae per atom.

These two one-dimensional hierarchies, of method and of basis set, can of course be combined. We will refer to any particular combination of an *ab initio* method, e.g., HF or MP2, and a generating basis set, e.g., cc-pVDZ or cc-pVTZ, as a *level of theory*. Each such pairing is viewed as providing an approximation  $V_{m,p} : (\mathbb{R}^3 \times \mathbb{N})^M \rightarrow \mathbb{R}$  to the true Born-Oppenheimer potential (2), where  $1 \leq m \leq 2N-1$  indexes the method as per the ordering of (3), and  $1 \leq p < \infty$  indexes the basis set within its particular family. In the last, it suffices intuitively to think of the cc-pVnZ basis sets, such that  $p=1$  indexes cc-pVDZ,  $p=2$  similarly cc-pVTZ, and so on, but any naturally-orderable family that produces a sufficiently regular error decay will do.

Similar two-dimensional hierarchies are widely used in *composite methods* like G4(MP2) [60, 61], the ccCA [62], W4 [63], and the HEAT schemes [64]; see also the reviews [7, 8]. As a general rule, these require some set of full-system single-point calculations to be carried out, each at a different level of theory. The energetic results of some of these are carefully subtracted from those of others, providing a collection of correction terms that represent, e.g., the added accuracy gained by moving from MP2/cc-pVDZ to CCSD/cc-pVDZ, or from MP2/cc-pVDZ to MP2/cc-pVTZ. The sum of these correction terms then gives an overall extrapolation towards a true FCI/CBS solution.

It was observed in passing in [14] that the derivation and resummation of the correction terms in composite methods is formally reminiscent of the *combination technique* [36, 37] for efficient high-dimensional approximation. In [6], this similarity was exploited to construct a *combination quantum machine-learning* scheme. For intuition, it helps to interpret the indices  $m$  and  $p$  as locating the potentials  $V_{m,p}$  on a so-called Pople diagram; see and cf. the reviews already cited, [5, Fig. 5.4], and [6, Fig. 1]. Without involving machine learning, a generalized picture of the connection can be given as follows, cf. [6, 8], although at some risk of oversimplification.<sup>2</sup> Define a family of contribution/correction terms by

$$\tilde{V}_{m,p} := V_{m,p} - \sum_{n < m} \tilde{V}_{n,p} - \sum_{q < p} \tilde{V}_{m,q} - \sum_{\substack{n < m \\ q < p}} \tilde{V}_{n,q}.$$

<sup>1</sup>Where we assume for the sake of simplicity that up to  $N$  spin orbitals are available for excitation. In practical settings, we should more properly consider the implications of, e.g., spin-restricted formalisms like RHF, see, e.g., [1, 3].

<sup>2</sup>For example, this does not allow for some common characteristics of composite models, such as the prevalent use of CBS extrapolation schemes, and the introduction of corrections targeting, e.g., core-valence interactions and relativistic effects. For a much more detailed discussion of this setup, see [47, Chap. 4].

Let  $\mathcal{L}$  be a linear functional mapping potentials  $V_{m,p}$  into some suitable Banach space  $Y$ , and assume much as in [14] that  $\|\mathcal{L}[V_{m,p} - V_{m,p-1}]\|_Y \lesssim g_p$ , for some sequence  $(g_p \in \mathbb{R})_{p \in \mathbb{N}}$  such that  $\sum_{p \in \mathbb{N}} g_p$  converges absolutely. We restrict ourselves here to a simple point-evaluation functional  $\mathcal{L}[V_{m,p}] = V_{m,p}(X_1, \dots, X_M)$  parametrized by a fixed molecular conformation  $\{X_A\}_{A=1}^M$ , but, e.g., an evaluation of the nuclear gradient requires no significant theoretical adjustment in what follows. Then we obtain pointwise an exact expansion of the Born-Oppenheimer potential as

$$V^{\text{BO}} = \sum_{m=1}^{2N-1} \sum_{p=1}^{\infty} \tilde{V}_{m,p}.$$

Experience with the combination technique suggests that accurate yet affordable results can be obtained by truncating this sum in a lower-triangular fashion, for example, after all terms with  $m+p \leq N'$  for some reasonably low  $N'$ . And indeed, up to very many intricacies, and precise choice of the set of potentials  $V_{m,p}$  and truncation of (2.1), this is essentially just how composite methods operate.

## 2.2. Energy-based fragmentation methods

Alternatively, *energy-based fragmentation methods*, hereafter just “fragmentation methods”, aim to obtain approximate solutions to the Schrödinger equation at reduced computational costs compared to the standard full-system methods such as those outlined above. For detailed reviews, see, e.g., [7, 9–11]. In brief, a fragmentation method involves a decomposition of the full set of nuclear indices  $[M] = \{1, 2, \dots, M\}$  into a family of  $K$  subsets  $F = \{F_i \subseteq [M]\}_{i=1}^K$  such that  $\bigcup_{i=1}^K F_i = [M]$ . As per, e.g., [7, 12, 13], this decomposition can be either *disjoint*, so that  $F_i \cap F_j = \emptyset$  for  $i \neq j$  and thus  $\{F_i\}_{i=1}^K$  is a strict partition of  $[M]$ , or *overlapping* if not; cf., however, comments in [10] questioning the meaningfulness of this differentiation in practice. We will refer to each  $F_i$  as a *fragment*, and the family  $F = \{F_i \subseteq [M]\}_{i=1}^K$  as a *fragmentation*, with the disjointness condition  $F_i \cap F_j = \emptyset$  holding unless otherwise stated.

The prototypical fragmentation method involves a truncation of the well-known *many-body expansion* (MBE), see, e.g., [13, 14, 33]. We construct the MBE as follows. Write simply as  $V : (\mathbb{R}^3 \times \mathbb{N})^M \rightarrow \mathbb{R}$  some particular symmetric potential, either an approximation  $V_{m,p}$  as defined above, or the true  $V^{\text{BO}}$ . Further, let  $\{V_{\mathbf{u}} : (\mathbb{R}^3 \times \mathbb{Z})^M \rightarrow \mathbb{R}\}_{\mathbf{u} \subseteq [M]}$  be a family of *subproblem potentials*, one for each subset of the nuclear indices  $[M]$ . Each  $V_{\mathbf{u}}$  is a potential which focuses basically the same level of theory as  $V$  on the nuclei indexed by  $\mathbf{u}$ ; however, each such potential still retains at least the possibility of a dependence on the full set of nuclei. This allows, for instance, the definition of each  $V_{\mathbf{u}}$  to cater for the introduction of such link atoms as are necessary after severing the subsystem indexed by  $\mathbf{u}$  out from its surrounding electrostatic environment [65], and/or to embed a single-point calculation in a field of Coulomb point charges, as in, e.g., [23, 66]. The MBE of  $V$  is then

$$V = \sum_{\mathbf{u} \subseteq [M]} \tilde{V}_{\mathbf{u}} = \tilde{V}_{\emptyset} + \sum_{i=1}^M \tilde{V}_{\{i\}} + \sum_{i=1}^M \sum_{j=i+1}^M \tilde{V}_{\{i,j\}} + \dots + \tilde{V}_{[M]}, \quad (4)$$

where each member of the family  $\{\tilde{V}_{\mathbf{u}}\}_{\mathbf{u} \subseteq [M]}$  is a *contribution potential*, defined for the moment recursively as

$$\tilde{V}_{\mathbf{u}} := V_{\mathbf{u}} - \sum_{\mathbf{v} \subset \mathbf{u}} \tilde{V}_{\mathbf{v}}. \quad (5)$$

As noted in [14], it is easy to see that as long as  $V_{[M]} = V$ , the expansion (4) is exact, regardless of the form of the remaining potentials  $V_{\mathbf{u}}$ . Although  $V_{\emptyset}$  can be zero, and usually implicitly is, this is not a requirement, as we shall briefly discuss below.

It is more common in practice to construct an MBE in terms of some fragmentation  $F$  of  $[M]$ , as in, e.g., [29]. To formulate this, writing  $F_{\mathbf{u}} := \bigcup_{i \in \mathbf{u}} F_i$ , we decompose  $V$  instead as

$$V = \sum_{\mathbf{u} \subseteq [K]} \tilde{V}_{F_{\mathbf{u}}}, \quad \tilde{V}_{F_{\mathbf{u}}} := V_{F_{\mathbf{u}}} - \sum_{\mathbf{v} \subset \mathbf{u}} \tilde{V}_{F_{\mathbf{v}}}. \quad (6)$$

We will refer to this form as a *fragment* MBE, in contrast to the *nuclear* MBE (4). For notational ease, we will work mostly with the nuclear form. But as we shall see, a fragment MBE can be viewed quite precisely as a restriction of a nuclear MBE.

The traditional way to derive a working fragmentation method from an MBE like (4) (or (6), proceeding equivalently) is to truncate the expansion after all terms  $\tilde{V}_{\mathbf{u}}$  with  $|\mathbf{u}| \leq n \leq M$  for some particular choice of  $n$ . The critical assumption driving such an *n-body expansion* is that of a decay in  $|\tilde{V}_{\mathbf{u}}|$  as  $|\mathbf{u}|$  increases, one swift enough that  $\sum_{|\mathbf{u}| > n} \tilde{V}_{\mathbf{u}}$  is negligibly small even for  $n \ll M$  [11, 14]. Then, since the evaluation costs of the potentials scale polynomially in  $|\mathbf{u}|$ , the terms in the *n-body expansion* are individually cheap to calculate, and this can collectively be done in an embarrassingly-parallel fashion [11]. It has been demonstrated in [28, 67] that these assumptions are flawed. The actual decay in the potentials is not as reliable or as fast as hoped, see also, e.g., [68], but more concerningly, the arithmetic required to evaluate an *n-body expansion* rapidly amplifies the uncertainties produced by the iterative solvers which always lie behind the potentials  $V_{\mathbf{u}}$  in practical implementation. We will briefly return to this latter point in Sections 3 and 4 below.

Other fragmentation methods start with an explicitly overlapping set of fragments, and generate weighted sums of energy terms motivated by inclusion/exclusion arguments in an effort to somehow avoid double-counting interactions or particles; see [13, 15], and cf., e.g., [17]. The *generalized many-body expansion* (GMBE) [13] seeks to build by extension a framework capable of handling *n-body* combinations of overlapping fragments, rather than disjoint fragments as in (6). To obtain an *n-body* truncation of the GMBE, one starts by forming the family  $\{F'_i\}_{i=1}^{K'}$  of all  $K' = \binom{K}{n}$  exactly *n-fold* unions of a set of  $K$  distinct and potentially overlapping fragments  $F_i$ , called in context *n-mers* [13, 15]. Adapting from [15, (1.9)], an *n-body* truncation of the GMBE can then be written as

$$E_{(n)}^{\text{GMBE}} = \sum_{i=1}^{K'} E_{F'_i} - \sum_{i < j}^{K'} E_{F'_i \cap F'_j} + \sum_{i < j < k}^{K'} E_{F'_i \cap F'_j \cap F'_k} + \cdots + (-1)^{K'+1} E_{F'_1 \cap \cdots \cap F'_{K'}}, \quad (7)$$

where, e.g.,  $E_{F'_i \cap F'_j} := \mathcal{L}[V_{F'_i \cap F'_j}]$  indicates the total energy of the subsystem indexed by the intersection of *n-mer* fragments  $F'_i$  and  $F'_j$ . The formula is explicitly motivated by the well-known inclusion/exclusion expression for the cardinality of the union of a collection of potentially-overlapping sets. In the case when the initial fragmentation is disjoint, the *n-body* GMBE energy is expected to reduce to a standard *n-body* truncation of a fragment MBE [13, 15].

Mathematically, the MBE is an *ANOVA-like decomposition* [14, 18, 19, 43, 44], and so each term  $\tilde{V}_{\mathbf{u}}$  can be viewed as the contribution of the lower-dimensional subset of dimensions indexed by  $\mathbf{u}$  to the high-dimensional function  $V$ . Just the same idea underlies a construction of the MBE as a particular *high-dimensional model representation* (HDMR) [45]. A low-order truncation of the MBE therefore exploits a low effective dimension in the underlying electronic problem [14, 43]. It has been widely observed, see, e.g., [7, 10, 12, 29], that the energy expressions of many fragmentation methods can be understood as truncated MBEs. We believe, probably uncontroversially, that the ANOVA-like viewpoint suggests that *any* fragmentation method should ultimately revert to a truncation  $\sum_{\mathbf{u} \in I} \tilde{V}_{\mathbf{u}}$  for a downward-closed subset  $I \subseteq 2^{[M]}$ , in the sense that if  $\mathbf{v} \subseteq \mathbf{u} \in I$  then also  $\mathbf{v} \in I$  — for if not, the “missing” terms in  $I$  introduce an inherent error to the approximation that cannot be systematically removed. This requirement has already been explicitly stated in some methods, e.g., [14, 29].

The following straightforward observation underpins much of what we do in the following: any truncation of (4) can be rewritten as a linear combination of potentials  $V_{\mathbf{u}}$  rather than explicitly contributions  $\tilde{V}_{\mathbf{u}}$ , simply by application of (5) and collection of terms. Doing so leads to non-recursive expressions for, e.g., standard *n-body* expansions [28, 69]. We highlight one particular existing setup which already acknowledges and exploits this for more arbitrary truncations. This *fragment combination range* (FCR) method [27, 29] is derived from a fragment MBE like (6). Adapting notation from [29] with reference also to [27, (9), (10), (11)], the total energy of a molecular system composed of  $K$  disjoint subsystems represented by composite coordinates  $\{z_i\}_{i=1}^K$  is approximated in the FCR approach as

$$E(z_1, \dots, z_K) \approx \sum_{\mathbf{f}_i \in \{\text{FCR}\}} p_{\mathbf{f}_i}^{\text{FCR}} E_{\mathbf{f}_i}(\{z\}_{\mathbf{f}_i}). \quad (8)$$

Here,  $\{\text{FCR}\}$  indicates an arbitrary downward-closed subset of  $2^{[K]}$ , and  $E_{\mathbf{f}_i}(\{z\}_{\mathbf{f}_i})$  the total energy of

the combined subsystem specified by those composite coordinates  $\{z_i\}_{i \in \mathbf{f}_l}$ . The coefficients  $p_{\mathbf{f}_l}^{\text{FCR}}$  are given non-recursively by

$$p_{\mathbf{f}_l}^{\text{FCR}} = \sum_{\substack{\mathbf{f}_{l'} \supseteq \mathbf{f}_l \\ \mathbf{f}_{l'} \in \{\text{FCR}\}}} (-1)^{l' - l}, \quad (9)$$

derived in [29] by an inductive argument. It is noted that in some cases, these coefficients vanish.

### 2.3. Fragmentation methods from the perspective of Möbius inversion

The principle of inclusion/exclusion, widely used to derive fragmentation method energy formulae like (7), can be substantially generalized by the technique of *Möbius inversion* from combinatorics and the theory of partially ordered sets [30–32]. For the convenience of the reader, we will reintroduce some basic ideas, following particularly [32] with minor notational deviation and general reference also to [30, 31]. We assume familiarity with the notion of a partially ordered set, or *poset*. A poset  $P$  is *locally finite* if  $\{u \in P \mid s \leq u \leq t\}$  is finite for every  $s \leq t$ .

**Definition 2.1** (Möbius function). The *Möbius function* of a locally finite poset  $P$  is

$$\mu_P(s, t) = \begin{cases} 1 & \text{if } s = t, \\ -\sum_{s \leq u < t} \mu(s, u) & \text{if } s < t, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

This is essentially the definition in [32], but extended to explicitly include the case when  $s \not\leq t$ . For particular posets  $P$ , simpler expressions exist for  $\mu_P$ . For example, suppose that  $P$  is a so-called *chain* poset, that is, a totally ordered set, and that it has a  $\hat{0}$ , i.e., a unique minimal element. Then  $P$  is isomorphic to  $\mathbb{N}$ , in the sense that there exists a bijection  $\phi : P \rightarrow \mathbb{N}$  which preserves  $\leq$ , as does its inverse  $\phi^{-1}$ . In this case, it is immediate from (10) that

$$\mu_P(s, t) = \begin{cases} 1 & \text{if } s = t, \\ -1 & \text{if } s \prec t, \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

where  $s \prec t$  indicates that  $s$  is *covered* by  $t$ , i.e.,  $s < t$  and there exists no intervening  $u \in P$  with  $s < u < t$ . Since the Möbius function is invariant under isomorphism, see also [70, Cors. 2 and 3], the form of  $\mu_{\mathbb{N}}$  follows immediately. Further, write  $B_n$  to be the powerset  $2^{[n]}$  ordered such that  $\mathbf{u} \leq \mathbf{v}$  exactly when  $\mathbf{u} \subseteq \mathbf{v}$ . This is the *Boolean algebra* of rank  $n$ , and (11) can be used to show that

$$\mu_{B_n}(\mathbf{u}, \mathbf{v}) = \begin{cases} (-1)^{|\mathbf{v} - \mathbf{u}|} & \text{if } \mathbf{u} \leq \mathbf{v}, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

The next result is a slightly specialized form of [32, Prop. 3.7.1], given without proof; historically, see [30]. We recall that the *principal order ideal* of some  $t \in P$  is the set  $\Lambda_t = \{s \in P \mid s \leq t\}$ .

**Theorem 2.1** (Möbius inversion [30, 32]). *If  $P$  is a poset such that  $\Lambda_t$  is finite for every  $t \in P$ , and  $f, g : P \rightarrow \mathbb{R}$ , then*

$$g(t) = \sum_{s \leq t} f(s)$$

*for all  $t \in P$  if and only if*

$$f(t) = \sum_{s \leq t} \mu(s, t) g(s)$$

*for all  $t \in P$ .*

A non-recursive expression for the terms  $\tilde{V}_{\mathbf{u}}$  in an MBE like (4) can be obtained pointwise via Möbius inversion [12, 33]. To make the bidirectional nature of the process clear, note that we can equally well just define outright each

$$\tilde{V}_{\mathbf{u}} := \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{v}-\mathbf{u}|} V_{\mathbf{v}} = \sum_{\mathbf{v} \subseteq \mathbf{u}} \mu_{B_M}(\mathbf{v}, \mathbf{u}) V_{\mathbf{v}}, \quad (13)$$

and then obtain (5) as an immediate consequence of Theorem 2.1. The equivalence between (5) and (13) is known even without explicit recourse to Möbius inversion [29], and has also been noted in the ANOVA setting; cf. [44, Thm. 2.1]. Interestingly, a simpler variant of Theorem 2.1 is invoked for this purpose in [17], where it is named as an extension of the PIE; cf. again [32].

We introduce some notation and ideas now in a very general form. These can be immediately specialized to the case of  $B_M$  currently under consideration, and therefore the MBE, but they will also apply to the general expansion form (23) to which we will progressively build. In combinatorial terminology, an *order ideal* is a subset  $I$  of a poset  $P$  such that if  $t \in I$ , then also  $s \in I$  for all  $s \leq t$ . Obviously, a downward-closed subset of  $B_M$  is just such an order ideal.

**Definition 2.2** (Combination coefficients). Let  $P$  be a locally finite poset with a  $\hat{0}$ , and let  $I$  be a finite order ideal of  $P$ . The *combination coefficient* of each  $s \in P$  for  $I$  is defined to be

$$D_s^{(I)} = \sum_{\substack{t \in I \\ t \geq s}} \mu_P(s, t). \quad (14)$$

The name “combination coefficient” stems from the combination technique, and there are direct connections in what follows to poset-based constructions in that area; see, e.g., [39–42] and the more general development in [47, Chap. 3]. From our perspective, the combination coefficients arise by rewriting what we will call an *I-truncation* of the MBE as a linear combination of the subproblem potentials  $V_{\mathbf{u}}$ , i.e.,

$$\begin{aligned} S_I &:= \sum_{\mathbf{u} \in I} \tilde{V}_{\mathbf{u}} = \sum_{\mathbf{u} \in I} \sum_{\mathbf{v} \subseteq \mathbf{u}} V_{\mathbf{v}} \mu_{B_M}(\mathbf{v}, \mathbf{u}) \\ &= \sum_{\mathbf{u} \in I} V_{\mathbf{u}} \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I}} \mu_{B_M}(\mathbf{v}, \mathbf{u}) = \sum_{\mathbf{u} \in I} D_{\mathbf{u}}^{(I)} V_{\mathbf{u}}, \end{aligned} \quad (15)$$

by (13) and rearrangement; cf. the alternative proof of [32, Prop. 3.7.1]. Note that when  $\mu_{B_M}(\mathbf{v}, \mathbf{u}) = (-1)^{|\mathbf{v}-\mathbf{u}|}$  is inserted in (14), then  $D_{\mathbf{u}}^{(I)}$  is exactly the FCR coefficient  $p_{\mathbf{f}_I}^{\text{FCR}}$  in (9), up to notation. So our setup can be viewed as just rederiving the FCR energy expression (8) by application of heavy machinery.

Now suppose that  $\hat{F} \subseteq B_M$  is an arbitrary subposet of  $B_M$ , recalling in general that a *subposet* of some poset  $P$  is a subset  $Q \subseteq P$  equipped with an appropriate restriction of the same order relation  $\leq$  as  $P$ . Then we can decompose

$$V \approx \sum_{\mathbf{u} \in \hat{F}} \tilde{V}'_{\mathbf{u}}, \quad (16)$$

in terms of contribution potentials defined only for elements of  $\hat{F}$ ,

$$\tilde{V}'_{\mathbf{u}} := V_{\mathbf{u}} - \sum_{\mathbf{v} \leq_{\hat{F}} \mathbf{u}} \tilde{V}'_{\mathbf{v}} = \sum_{\mathbf{v} \leq_{\hat{F}} \mathbf{u}} \mu_F(\mathbf{v}, \mathbf{u}) V_{\mathbf{v}},$$

where the last equality follows again by Möbius inversion. For reasons that will become clear below, we will refer to (16) as a SUPANOVA expansion, but it can also be viewed as a special case of the CGTCE decomposition form given in [34]. If  $[M] \in \hat{F}$  and  $V_{[M]} = V$ , then the decomposition (16) is exact, but even if not, a sum  $S'_{I'} = \sum_{\mathbf{u} \in I'} \tilde{V}'_{\mathbf{u}}$  can be defined for any order ideal  $I'$  of  $\hat{F}$ , and rewritten as a linear combination of only those original contribution potentials  $V_{\mathbf{u}}$  where  $\mathbf{u} \in \hat{F}$ . Such a truncation  $S'_{I'}$  can be exactly identified with some  $S_I$  for an order ideal of the full  $B_M$  at least when the non-zero terms of  $S_I$  are exactly those of  $S'_{I'}$ . We can formalize this in the general setting as follows:

**Definition 2.3** (Combination-consistency). Let  $P$  be a locally finite poset, and  $Q$  be a subposet of  $P$ . Further, let  $I$  be a finite order ideal of  $P$ , and  $I'$  be a finite order ideal of  $Q$ . Write as above  $D_s^{(I)}$  to be the combination coefficient of  $s \in P$  for  $I$ , and for notational clarity, write  $\hat{D}_{s'}^{(I')}$  to be the combination coefficient of  $s' \in Q$  for  $I'$ . If for every  $s \in P$  it holds that

$$D_s^{(I)} = \begin{cases} \hat{D}_s^{(I')} & \text{if } s \in Q, \\ 0 & \text{otherwise,} \end{cases} \quad (17)$$

then  $I$  and  $I'$  are *combination-consistent*. If a combination-consistent finite order ideal  $I$  of  $P$  exists for every finite order ideal  $I'$  of  $Q$ , then  $Q$  is a *combination-consistent* subposet of  $P$ .

If we assume that  $P$  possesses a certain kind of structure, we can characterize the necessary requirements for some subposet  $Q$  of  $P$  to be combination-consistent with it. This characterization slightly generalizes work in [35]. For a detailed explanation of the connection, see [47, Sec. 5.2]. But in short, a cluster expansion is there truncated after a certain order ideal of a direct product of subposets of a Boolean algebra, and Möbius inversion leads to a different expression for what we call combination coefficients. This expression can be recreated in our setup as follows, as an immediate consequence of (14) and (10). A direct connection also exists here to a textbook derivation of the set-cardinality PIE expression via Möbius inversion; see and cf. [32, p. 265]. The PIE in turn connects back to the combination technique; see, e.g., [40, 42].

**Lemma 2.1.** *Let  $I$  be an order ideal of a locally finite poset  $P$ . Define  $J := I \cup \{\hat{1}_J\}$  ordered just as  $I$  but with additionally  $s < \hat{1}_J$  for every  $s \in I$ . The precise choice of  $\hat{1}_J$  is not important here, only that there exists some unique maximal element of  $J$ . Then, for every  $s \in I$ ,*

$$D_s^{(I)} = -\mu_J(s, \hat{1}_J).$$

We need some further basic definitions; see again and cf., e.g., [31, 32]. A *meet semilattice* is a poset  $P$  where every pair of elements  $s, t \in P$  have a unique greatest lower bound  $r$  in  $P$ ; this bound is called their *meet* and written as  $r = s \wedge t$ . If  $Q$  is a meet-closed subposet of a meet semilattice  $P$ , i.e., one where  $s \wedge_P t \in Q$  for any  $s, t \in Q$ , then we call  $Q$  a *meet subsemilattice* of  $P$ . For a finite order ideal  $I$  of a general poset  $P$ , the *generating antichain* of  $I$  is the set  $A = \{s \in I \mid \nexists t \in I, t > s\}$  of maximal elements of  $I$ ; we will write then  $I = \langle A \rangle := \bigcup_{a \in A} \Lambda_a$ . Using this terminology, we can rephrase a key observation made on [35, p. 7481] in a more general setting. The proof is given in Appendix A, and uses just the same ideas as in [35].

**Proposition 2.1** (Generalized from [35]). *Let  $P$  be a locally finite meet semilattice, and  $I$  be a finite order ideal of  $P$  with generating antichain  $A$ . If  $D_s^{(I)} \neq 0$  for some  $s \in P$ , then there exists some  $\{a_1, \dots, a_n\} \subseteq A$  such that  $s = a_1 \wedge \dots \wedge a_n$ .*

Using our notation and terminology, and with a little generalization, it was further observed in [35] that the subposet  $Q$  built from the meets of an essentially arbitrary antichain  $A$  of a particular meet semilattice  $P$  is combination-consistent with  $I = \langle A \rangle_P \subseteq P$ . It is straightforward to check that, given an order ideal  $I'$  of some arbitrary subposet  $Q$  of a more general  $P$ , the only possible combination-consistent order ideal  $I$  of  $P$  is that provided by the generating antichain of  $I'$ , if one exists at all. For the proof, see Appendix A.

**Lemma 2.2** (Adapted from Lem. 5.2.7 [47]). *Let  $P$  be a locally finite poset, let  $Q$  be a subposet of  $P$ , and let  $I' \subseteq Q$  be an arbitrary finite order ideal of  $Q$ . If  $I$  is a finite order ideal of  $P$  which is combination-consistent with  $I'$ , then  $I = \langle A' \rangle_P$ , where  $A' \subseteq I'$  is the generating antichain of  $I'$ .*

Since the next result is an immediate consequence of standard results in order theory, it is hardly interesting in its own right, although we do not believe we have seen it explicitly stated as such. But it is very important for us. The key ideas, particularly for direction ( $\Leftarrow$ ), are again to be found in [35] for a particular lattice, and their generalization to an arbitrary meet semilattice is immediate. The proof is relegated to Appendix A. We remark that other lattices, related to  $\mathbb{N}^d$ , also arise very naturally in the standard grid-based setting of the combination technique [39–42], and there is also an especially close connection here to an inclusion/exclusion-based construction in [42, Chap. 3]; see [47, Sec. 5.2.3] for deeper discussion.

**Theorem 2.2** (Thm. 5.2.8 [47]). *Let  $P$  be a locally finite meet semilattice with a  $\hat{0}$ , and let  $Q$  be a subposet of  $P$ . Then  $Q$  is combination-consistent with  $P$  if and only if  $Q$  is a meet subsemilattice of  $P$ .*

As a first consequence of this, consider some fragmentation  $F = \{F_i \subseteq [M]\}_{i=1}^K$  of  $[M]$ , and let in this case  $\hat{F} = \{F_{\mathbf{u}} \mid \mathbf{u} \in B_K\}$ . Since meets in  $B_M$  are given by set intersections, it is not hard to see that  $\hat{F}$  is a meet subsemilattice of  $B_M$ . This means that any downward-closed truncation of a fragment MBE (6), standard  $n$ -body or otherwise, can also be understood as a downward-closed truncation of a nuclear MBE (4). Note that while the former fragment-MBE truncation can then be exactly expressed as a sum of nuclear-MBE contributions  $\sum_{\mathbf{u} \in I} \tilde{V}_{\mathbf{u}}$  indexed by the appropriate  $I \subseteq B_M$ , only the nuclear-MBE subproblem potentials  $V_{\mathbf{u}}$  for  $\mathbf{u} \in \hat{F}$  actually matter in this sum, and the remainder indexed by  $B_M - \hat{F}$  can be chosen completely arbitrarily.

Theorem 2.2 applies to any arbitrary subposet  $\hat{F}$  of  $B_M$ . This can be used to rigorously understand some, and we claim most if not all, of the various fragmentation methods that build from sets of overlapping fragments just as producing particular truncations of a fragment MBE, and therefore transitively an underlying nuclear MBE. This has been informally anticipated, and in some cases either shown directly or argued by example, see again [7, 10, 29]. The following result is hypothesized in [29], but with only a non-rigorous justification. Some further practical demonstration is given in [27]. The proof is condensed and clarified from the version in [47, Sec. 5.2]. It should be stressed that this is, in essence, just a variant of the same argument used in [32, p. 265] to construct the set-cardinality PIE using Möbius inversion. Observe also the connection to the decomposition in terms of the meets of an antichain in [35].

**Proposition 2.2.** *Let  $F = \{F_i\}_{i=1}^K$  be a not necessarily disjoint fragmentation of  $[M]$ . Assuming that  $V_{\emptyset} = 0$ , then for any  $1 \leq n \leq K$  there exists an order ideal  $I$  of  $B_M$  such that  $S_I = E_{(n)}^{\text{GMBE}}$ .*

*Proof.* Let  $\{F'_i\}_{i=1}^{K'}$  be the set of  $n$ -mers required for an  $n$ -body GMBE, built from  $F$ . Notationally, define  $F'_{\cap \mathbf{u}} := \bigcap_{i \in \mathbf{u}} F'_i$  for every  $\emptyset \subset \mathbf{u} \subseteq [K']$ . Take  $\hat{F} := \{F'_{\cap \mathbf{u}} \mid \emptyset \subset \mathbf{u} \subseteq [K']\} \cup \{\emptyset\}$ , cf. [32, p. 265], and rewrite (7) as

$$E_{(n)}^{\text{GMBE}} = \sum_{\emptyset \subset \mathbf{u} \subseteq [K']} (-1)^{|\mathbf{u}|+1} E_{F'_{\cap \mathbf{u}}} = \sum_{\substack{\hat{\mathbf{u}} \in \hat{F} \\ \hat{\mathbf{u}} \neq \emptyset}} d_{\hat{\mathbf{u}}} E_{\hat{\mathbf{u}}}, \quad (18)$$

where we define  $d_{\hat{\mathbf{u}}}$  to be the sum of all  $\pm 1$  coefficients in terms  $(-1)^{|\mathbf{u}|+1} E_{F'_{\cap \mathbf{u}}}$  where  $F_{\cap \mathbf{u}} = \hat{\mathbf{u}}$ , cf., e.g., [66].

We will show that each  $d_{\hat{\mathbf{u}}} = D_{\hat{\mathbf{u}}}^{(\hat{F})}$ , where the latter is the combination coefficient of  $\hat{\mathbf{u}} \in \hat{F}$  viewed as an order ideal of itself. First, fix some non-empty  $\hat{\mathbf{u}} \in \hat{F}$ . By construction, there exists at least one  $\mathbf{u} \subseteq [K']$  such that  $F'_{\cap \mathbf{u}} = \hat{\mathbf{u}}$ ; moreover, there exists exactly one such  $\mathbf{u}$  with maximal  $|\mathbf{u}|$ , for if there were two or more,  $\mathbf{u}_1, \mathbf{u}_2, \dots$ , it would hold that  $\hat{\mathbf{u}} = F'_{\cap(\mathbf{u}_1 \cup \mathbf{u}_2)}$ , contradicting the maximality of  $|\mathbf{u}_1|$  and  $|\mathbf{u}_2|$ . Note then that for each  $\emptyset \subset \mathbf{v} \subseteq \mathbf{u}$ , it holds that  $F'_{\cap \mathbf{v}} = \hat{\mathbf{v}}$  for some  $\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}$ , and conversely, each  $\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}$  can be written as  $F'_{\cap \mathbf{v}}$  for at least one  $\emptyset \subset \mathbf{v} \subseteq \mathbf{u}$ . Therefore,

$$\sum_{\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}} d_{\hat{\mathbf{v}}} = \sum_{\emptyset \subset \mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{v}|+1} = \sum_{j=1}^{|\mathbf{u}|} (-1)^{j+1} \binom{|\mathbf{u}|}{j} = 1 - \sum_{j=0}^{|\mathbf{u}|} (-1)^j \binom{|\mathbf{u}|}{j} = 1, \quad (19)$$

where we use a standard combinatorial identity, namely [71, (1.31)]. Finally, by Möbius inversion of (19) considering the dual  $\hat{F}^*$  of  $\hat{F}$ , that is,  $\hat{F}$  reordered by  $\mathbf{u} \geq_{\hat{F}^*} \mathbf{v}$  iff  $\mathbf{u} \leq_{\hat{F}} \mathbf{v}$ , and according to (14),

$$d_{\hat{\mathbf{u}}} = \sum_{\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}} \mu_{\hat{F}^*}(\hat{\mathbf{v}}, \hat{\mathbf{u}}) = \sum_{\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}} \mu_{\hat{F}}(\hat{\mathbf{u}}, \hat{\mathbf{v}}) = D_{\hat{\mathbf{u}}}^{(\hat{F})}.$$

Since  $V_{\emptyset} = 0$ , then, it follows that (18) is just (16) for this particular  $\hat{F}$ . Since  $\hat{F}$  is closed under intersection by construction, it is a meet subsemilattice of  $B_M$ , and the desired result follows from Theorem 2.2 by choosing  $I' = \hat{F}$ .  $\square$

We observe that (19) coincides with a more general “top-down” expression for the FCR coefficients derived by a counting argument in [27], which in our notation is  $D_{\mathbf{u}}^{(I)} = 1 - \sum_{\mathbf{u} \subset \mathbf{v} \in I} D_{\mathbf{v}}^{(I)}$ . In fact, and

more generally again, given any finite order ideal  $I$  of a suitable poset  $P$ , dual-form Möbius inversion (see [32, Prop. 3.7.2]) of (14) immediately delivers  $D_s^{(I)} = 1 - \sum_{s < t \in I} D_t^{(I)}$  for arbitrary  $s \in I$ . This can be understood as a property of the Möbius function, cf. Exercise 3.88 of [32].

We required  $V_\emptyset = 0$  in the statement of Proposition 2.2 in order that the truncation  $S_I$  should match the GMBE form (7), and this is more generally the intuitively correct selection if each  $V_{\mathbf{u}}$  should provide the total energy of the subsystem indexed by  $\mathbf{u}$  held in isolation. However, in [66], the energy terms  $E_{\mathbf{u}}$  in what can be viewed as a one-body GMBE are calculated by embedding subsystems in a field of electrostatic point charges, and an overcounting correction for the self-interactions of those charges is added to (7), with form justified again via the PIE. This correction can be viewed as a non-zero expression for  $V_\emptyset$ , weighted by  $d_\emptyset = 1 - \sum_{\emptyset \neq \hat{\mathbf{u}} \in \hat{F}} d_{\hat{\mathbf{u}}}$  in the above notation, so a special case of (19) and the expressions in the previous paragraph.

This, then, is the logic behind our general definition of the subproblem potentials  $V_{\mathbf{u}} : (\mathbb{R}^3 \times \mathbb{Z})^M \rightarrow \mathbb{R}$ . If, e.g., a family of embedding-style potentials are constructed, such that  $V_\emptyset$  represents an extremely coarse model of the full system and each  $V_{\mathbf{u}}$  somehow improves the treatment of the subsystem indexed by  $\mathbf{u}$  compared to those by each  $\mathbf{v} \subset \mathbf{u}$ , then a kind of “overcounting correction” analogous to that in [66] is automatically built into any truncation by way of  $D_\emptyset^{(I)}$ . This also applies to truncations of the multilevel decomposition (23) below. The idea should find natural application to MBE-like expansions constructed using explicitly quantum embedding techniques; see the review [72] and cf., e.g., quite recent work in [73, 74]. Although we will not explicitly treat quantum-embedding potentials here, we considered some initial experiments in this direction in [47]; see there for details and discussion. Although these results were mixed, we still consider this a very promising area for future investigation.

## 2.4. Constructing consistent graph-based decompositions

Most fragmentation methods exploit either distance-based thresholding criteria or some more abstract idea of structural connectivity in order to exclude some energetic terms from calculation; see again the reviews [7, 9–11] for many examples. Implicitly or explicitly graph-theoretic techniques can be used to choose an initial fragmentation directly, and/or to derive a decomposition [14, 17–22, 75]. For graph-theoretic notation and terminology, we follow mostly [76], with general reference also to [77, 78].

Fundamental in this context is the representation of a molecular system via an *interaction graph*, that is, an undirected graph  $G = (V = [M], E)$ , with an edge set chosen to capture some picture of the connectivity of the individual atoms in the system. Although we have also used  $V$  to denote potentials and  $E$  energies in the above, the use of standard graph-theoretical notation should be unambiguous in context. Surely the canonical choice for an interaction graph is a *bond graph*, where each edge  $\{i, j\} \in E$  corresponds to a covalent bond between the two atoms with indices  $i, j$ . However, the bond graph is by no means the only plausible interaction graph; see, e.g., [17, 21, 79].

Once an interaction graph is obtained, subproblems can be identified from particular subgraphs of  $G$ . A full comparison of the many different ways this can be done is beyond the scope of this article. But for example, the BOSSANOVA decomposition [18, 19] is essentially a modified MBE that includes only contribution terms for those subsets  $\mathbf{u} \subseteq [M]$  which induce connected subgraphs  $G[\mathbf{u}]$  of  $G$ . We write the set of all such subsets as  $\text{conn}[G]$  and treat it as a subposet of  $B_M$ . The informal motivation, consistent with Kohn’s famous nearsightedness principle [80], is that at any given order  $1 \leq |\mathbf{u}| \leq M$ , these subgraphs should provide a tidy way to select the most important  $|\mathbf{u}|$ -body interactions in the system. Some potential  $V$  is decomposed as

$$V = \sum_{\mathbf{u} \in \text{conn}[G]} \tilde{V}_{\mathbf{u}}^{\text{BOSSANOVA}}, \quad \tilde{V}_{\mathbf{u}}^{\text{BOSSANOVA}} := V_{\mathbf{u}} - \sum_{\substack{\mathbf{v} \in \text{conn}[G] \\ \mathbf{v} \subset \mathbf{u}}} \tilde{V}_{\mathbf{v}}^{\text{BOSSANOVA}}. \quad (20)$$

Of course, (20) is just an instantiation of (16) for  $\hat{F} = \text{conn}[G]$ . As mentioned above, (16) and thus (20) can be recognized as special cases of the very general *chemical graph-theoretic cluster expansion* (CGTCE) [34]. Since the set of induced subgraphs of  $G$  ordered by their inducing sets is clearly isomorphic to  $B_M$  [46], it is only a matter of perspective to regard any decomposition (16) as being determined by a particular collection of induced subgraphs rather than by a particular subposet of  $B_M$ . However, we will refer to any explicitly induced-subgraph-based decomposition of the form (16) as a *SUPANOVA*



*decomposition*, for *SUBgraph Poset ANOVA*. We take the liberty of using a different name since, in context, we treat this setup primarily as a useful extension on the BOSSANOVA decomposition, and we will extend it below to generalize on the multilevel ML-BOSSANOVA scheme [14]. To our reading, such a multilevel adjustment to the CGTCE was not envisioned in [34].

Although the results of initial experiments with  $n$ -body truncations of the BOSSANOVA decomposition over chain-type molecules were highly encouraging [18, 19], difficulties arose for molecules containing ring substructures. These were initially ascribed to physical effects in ring fragmentation. While these are important, a more subtle, yet critical issue becomes obvious in light of Theorem 2.2.

It is rather easy to find graphs  $G$  where  $\text{conn}[G]$  is not closed under intersection and therefore cannot be a meet subsemilattice of the appropriate Boolean algebra. Consider, e.g., the cycle graph  $C_6 = (V = [6], E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{5, 6\}, \{6, 1\}\})$ , and note that the intersection of  $\{1, 2, 3, 4\}$  and  $\{1, 4, 5, 6\}$  does not induce a connected subgraph. This graph is isomorphic to the covalent bond graph of benzene, neglecting hydrogens. In general, for such a  $G$ , there must exist at least one, and practically there will be many order ideals of  $\text{conn}[G]$  that do not have a combination-consistent order ideal of  $B_M$ . The corresponding truncations of (20) effectively miscount MBE contribution terms  $\tilde{V}_{\mathbf{u}}$ ; for a full worked example demonstrating this effect, see [47, Sec. 6.3]. Interestingly, this seems to be the same issue which necessitates the “ring repair rule” in the SMF method [75], cf. [81], although we do not attempt to argue this rigorously.

The conditions on  $G$  under which  $\text{conn}[G]$  is a sublattice of  $B_M$  were established in [46]. It follows in particular from [46, Lems. 1 and 2] that  $\text{conn}[G]$  cannot be combination-consistent with  $B_M$  if  $G$  contains certain forbidden induced subgraphs, specifically any chordless cycle of length greater than three, or a cycle of length four with one chord. For details and more precise formulation, see again [47, Sec. 6.3]. This rules out the useful application of the BOSSANOVA decomposition (20) in terms of, for instance, the bond graphs of the vast majority of cyclic compounds.

We seek, then, a suitable subposet of subgraphs of  $G$  that embody a similar kind of locality to the connected subgraphs, while still being a combination-consistent subposet of  $B_M$ . One possibility is to use the geodesically convex subgraphs of  $G$ . In fact, these were mentioned on [34, p. 156] as providing a “well-behaved” option in a particular application of the CGTCE, namely the construction of a certain coupled cluster-like expansion. We give here only a self-contained definition of the geodesically convex subgraphs of a graph that is essentially the same as that used there. A deeper understanding of these and other kinds of convex subgraphs should be grounded in the theory of abstract convexity; see, e.g., [82] for general background, [83] for a treatment of graph convexities in particular, and [47, Sec. 6.5] for more detailed summary and further references. We base our notation on that in, e.g., [84].

**Definition 2.4** (Geodesically convex subgraphs of an interaction graph). Let  $G$  be a connected interaction graph, with vertex set  $V = [M]$ . An induced subgraph  $G[\mathbf{u}]$  of  $G$  is (*geodesically*) *convex* if, for any pair of vertices  $i, j \in \mathbf{u}$ , all vertices along any shortest path between  $i$  and  $j$  in the full graph  $G$  lie also in  $G[\mathbf{u}]$ . The *geodesic convexity* of  $G$ , written  $\mathcal{M}_g[G]$ , is the subposet of  $B_M$  formed by all  $\mathbf{u} \in B_M$  such that  $G[\mathbf{u}]$  is a convex subgraph of  $G$ .

Combination-consistency of  $\mathcal{M}_g[G]$  with  $B_M$  results from the obvious fact that if  $G[\mathbf{u}]$  and  $G[\mathbf{v}]$  are convex subgraphs of  $G$ , then so too is  $G[\mathbf{u} \cap \mathbf{v}]$ . However,  $\mathcal{M}_g[G]$  is not always an ideal poset for use in a SUPANOVA expansion (16). An important result [85, Thm. 4.1] shows that  $\mathcal{M}_g[G]$  fails in many cases of practical interest to be a *convex geometry* [82]. In particular, and informally, subsets  $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{v}$  may then have substantially different sizes  $|\mathbf{u}|$  and  $|\mathbf{v}|$ ; cf. [82, Thm. 2.1]. This has practical implications for the adaptive algorithm we give below. It would be very interesting to investigate whether general schemes could be developed for the construction and/or adjustment of  $G$  in a way that would ensure that  $\mathcal{M}_g[G]$  is a convex geometry, but this remains for future work.

There is an interesting connection between a *convex SUPANOVA* decomposition, that is, (16) for  $\mathcal{M}_g[G]$ , and the energy expression of a multilevel graph-based fragmentation method described in, e.g., [20, 21]. We consider explicitly only a non-multilevel expression given in [21]. In the language of the scheme, if an induced subgraph  $G[\mathbf{u}]$  of  $G$  is complete, it is called a *rank- $r$  simplex*, where  $r = |\mathbf{u}| - 1$ . We write  $\text{comp}_R[G]$  to collect those subsets of  $[M]$  which induce simplices of rank up to some  $0 \leq R \leq M - 1$ .

Then the energy expression is, adapting [21, (3)],

$$E_R^{\text{simplex}} = \sum_{r=0}^R (-1)^r \sum_{\substack{\mathbf{u} \in \text{comp}_R[G] \\ |\mathbf{u}|=r+1}} E_{\mathbf{u}} \left( \sum_{m=r}^R (-1)^m p_{\mathbf{u}}^m \right), \quad (21)$$

where  $p_{\mathbf{u}}^m$  counts the number of rank- $m$  simplices which include the rank- $r$  simplex  $G[\mathbf{u}]$  as a subgraph.

As noted for simplices in, e.g., [20], if  $\mathbf{u} \subseteq \mathbf{v}$ , and  $G[\mathbf{v}]$  is complete, then so too is  $G[\mathbf{u}]$ . As a result,  $\text{comp}_R[G]$  is an order ideal of  $B_M$ . The expression (21) can be obtained directly by manipulation of (15) after a corresponding insertion of (12); indeed, this is just a special case of an expression already given in the FCR setting, see [29, (14), (15)]. So (21) is simply a particular truncation of (4); this was recognized in, e.g., [20, 21]. But note also that every complete induced subgraph of  $G$  is also a convex subgraph, although not necessarily vice-versa. As a result,  $\text{comp}_R[G]$  is also an order ideal of  $\mathcal{M}_g[G]$ , and so (21) is one truncation of a convex SUPANOVA decomposition.

It is not hard to imagine situations where it would be important to have access to the full range of convex SUPANOVA truncations, rather than only the subset provided by (21). Consider, for instance, a chain system of disjoint fragments, for instance, a linear alkane. Here, the only simplices would be the empty graph ( $R = -1$ ), the individual fragments ( $R = 0$ ), and neighboring pairs ( $R = 1$ ). The full graph  $G$  is not complete, so  $E_R^{\text{simplex}}$  cannot “converge” to the full-system energy as  $R$  increases, and, e.g., neighboring triples of fragments are not simplices, so the descriptive power of the simplex approach for longer-range interactions is inherently limited to the edge set of  $G$ . Were a full convex SUPANOVA expansion considered instead, the adaptive method for truncation selection that we outline below should be able to explore up through the additional terms, and also obviate the need to explicitly preselect a maximum rank  $R$  as a parameter.

## 2.5. Combining composite and fragmentation methods

Although the true Born-Oppenheimer potential  $V^{\text{BO}}$  can in principle be expanded as, e.g., an MBE like (4), this is hardly useful, since the potentials  $V_{\mathbf{u}}$  still remain numerically inaccessible. Practically, instead, one must consider an expansion instead of a full-system potential  $V_{m,p}$  at a single level of theory. Multilevel fragmentation methods [14, 20, 21, 23–27] attempt to achieve an effective approximation of such a full-system potential at a high-quality level of theory via evaluations of some of the subproblem potentials  $V_{\mathbf{u}}$  at the same high level, and of others at a lower and thus computationally cheaper level. In this section, we interlace ideas from the previous sections with those found in existing multilevel methods in order to derive a new and very general multilevel fragmentation approach.

By *multilevel*, we refer most generally to any scheme which produces a single approximate solution to (1) from a collection of other and presumably less-accurate solutions calculated from multiple levels of theory. The composite methods mentioned in Section 2.1 clearly form a class of multilevel approach [6]. Another, very well-known such class is provided by the multilayered ONIOM method [86] and its many variants and derivatives [87]. Stated most simply, the principle of ONIOM is to approximate the total energy  $E^{\text{HL}}$  of a molecular system as per some high level of theory as

$$E^{\text{HL}} \approx E^{\text{LL}} + (E_{\mathbf{u}}^{\text{HL}} - E_{\mathbf{u}}^{\text{LL}}),$$

see and cf. [86, (3)]. Here,  $E^{\text{LL}}$  is the full-system total energy calculated instead using a lower level of theory, and  $E_{\mathbf{u}}^{\text{HL}}$  and  $E_{\mathbf{u}}^{\text{LL}}$  the high/low-level total energies of a subsystem indexed by  $\mathbf{u} \subset [M]$  and deserving of accentuation. The idea extends to a successively-nested chain of  $n$  subsystems  $[M] = \mathbf{u}_1 \supset \mathbf{u}_2 \supset \cdots \supset \mathbf{u}_n$ , each matched with an increasingly stronger level of theory; see and cf. [86, (4)].

The core ONIOM idea has provided a springboard for the development of many multilevel fragmentation methods; see, e.g., [20–22, 26, 88], and also again [7, 9, 11, 87]. To connect these and other multilevel fragmentation methods with composite methods, we introduce a very general expansion of some potential  $V : (\mathbb{R}^3 \times \mathbb{N})^M \rightarrow \mathbb{R}$ . We begin by recalling, see, e.g., [32], that the *direct product* of two posets  $P \times Q$  is ordered such that  $(s, t) \leq_{P \times Q} (s', t')$  iff  $s \leq_P s'$  and  $t \leq_Q t'$ . The following basic result is very helpful; we omit the proof.

**Theorem 2.3** (Product theorem [30, 32]). *For the direct product  $P \times Q$  of two locally finite posets  $P$  and  $Q$ , it holds that*

$$\mu_{P \times Q}((s, t), (s', t')) = \mu_P(s, s') \mu_Q(t, t'). \quad (22)$$

Call the members of some family of locally finite posets  $\{P_i\}_{i=1}^d$  to be poset *axes*, and by extension their direct product  $\Pi = P_1 \times \cdots \times P_d$  to be a  $d$ -dimensional poset *grid*. Note that  $\mu_\Pi$  can be obtained from (22) by recursively rewriting, e.g.,  $\Pi = P_1 \times (P_2 \times \cdots \times P_d)$ . The axes  $P_i$  need not be finite, but we will assume them to each have a  $\hat{0}$ . Then, analogously to above, we assume the existence of some family of subproblem potentials  $\{V_{\mathbf{p}}\}_{\mathbf{p} \in \Pi}$ , and expand some  $V : (\mathbb{R}^3 \times \mathbb{N})^M \rightarrow \mathbb{R}$  as

$$V = \sum_{\mathbf{p} \in \Pi} \tilde{V}_{\mathbf{p}}, \quad \tilde{V}_{\mathbf{p}} := V_{\mathbf{p}} - \sum_{\mathbf{q} < \mathbf{p}} \tilde{V}_{\mathbf{q}} = \sum_{\mathbf{q} \leq \mathbf{p}} \mu_\Pi(\mathbf{q}, \mathbf{p}) V_{\mathbf{q}}. \quad (23)$$

This equality, and convergence of the sum in the case of one or more infinite axes  $P_i$ , is conditional on the precise choice of subproblem potentials. The idea here is to choose these and also the overall grid  $\Pi$  in such a way that pointwise convergence to  $V^{\text{BO}}$  holds, practically if not provably. In any case, given any finite order ideal  $I$  of  $\Pi$ , the  $I$ -truncation  $S_I = \sum_{\mathbf{p} \in I} \tilde{V}_{\mathbf{p}}$  certainly exists, and can be converted into a sum in terms of combination coefficients  $D_{\mathbf{p}}^{(I)}$  and subproblem potentials  $V_{\mathbf{p}}$ .

For particular choices of  $\Pi$ , the expansion form (23) captures not only the nuclear MBE (4), the fragment MBE (6), the BOSSANOVA decomposition (20) and its SUPANOVA (or CGTCE) generalization (16), but also, for  $\Pi = [2N - 1] \times \mathbb{N}$ , the composite-style expansion (2.1). It can also reproduce a number of existing multilevel fragmentation methods. In the interest of space, we will only explicitly show this for the relatively recent example of the multilevel ML-FCR approach [27], which is itself an extension on the original FCR method. Note that, in the case of  $B_M$ , the FCR expression (8) was completely equivalent to an  $I$ -truncation  $S_I$ , and the benefit of our approach was in the supporting theoretical tools provided. In the multilevel setting, however, although the ML-FCR is already very general, we can extend it further.

Only a two-level ML-FCR formulation is explicitly given in [27]. An extension to further layers is suggested but not given. Such an extension seems to be in use in [22], but we restrict ourselves here to the original formulation. Two separate downward-closed sets of potentially-overlapping fragments are considered, which we write as  $\{\text{FCR}_{\text{HL}}\}$  and  $\{\text{FCR}_{\text{LL}}\}$ . Generally, one would expect the former to be a subset of the latter, but the involved formulae work correctly even when this relationship is inverted. Adjusting notation to match (8) above, the ML-FCR energy equation is

$$E^{\text{ML-FCR}}(z_1, \dots, z_K) = \sum_{\mathbf{f}_l \in \{\text{FCR}_{\text{HL}}\}} p_{\mathbf{f}_l}^{\text{HL}} E_{\mathbf{f}_l}^{\text{HL}}(\{z\}_{\mathbf{f}_l}) + \sum_{\mathbf{f}_l \in \{\text{FCR}_{\text{LL}}\}} p_{\mathbf{f}_l}^{\text{LL}} E_{\mathbf{f}_l}^{\text{LL}}(\{z\}_{\mathbf{f}_l}), \quad (24)$$

see and cf. [27, (15)]. Each energy term  $E_{\mathbf{f}_l}^{\text{HL}}$  is calculated with a somehow high-level method, and each  $E_{\mathbf{f}_l}^{\text{LL}}$  with a low-level method. The high-level coefficients  $p_{\mathbf{f}_l}^{\text{HL}}$  are non-multilevel FCR coefficients (9) for  $\mathbf{f}_l$  in  $\{\text{FCR}_{\text{HL}}\}$ . The high- and low-level coefficients are connected by

$$p_{\mathbf{f}_l}^{\text{HL}} + p_{\mathbf{f}_l}^{\text{LL}} = p_{\mathbf{f}_l}^{\{\text{FCR}_{\text{HL}}\} \cup \{\text{FCR}_{\text{LL}}\}}, \quad (25)$$

motivated in [27] via a counting argument, where similarly  $p_{\mathbf{f}_l}^{\{\text{FCR}_{\text{HL}}\} \cup \{\text{FCR}_{\text{LL}}\}}$  is the non-multilevel FCR coefficient for  $\mathbf{f}_l$  in  $\{\text{FCR}_{\text{HL}}\} \cup \{\text{FCR}_{\text{LL}}\}$ .

To reconstruct the ML-FCR from (23), including a generalization to multiple layers, we choose  $\Pi = B_M \times [n]$ , for some  $n \geq 1$ . For an order ideal  $I$  of  $\Pi$ , we can write  $I = \bigcup_{i=1}^n I_i \times \{i\}$ , where each  $I_i$  is an order ideal of  $B_M$  and also  $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n$ . Recalling (14) and using (22), we have that

$$D_{(\mathbf{u}, n)}^{(I)} = \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_n}} \mu_\Pi((\mathbf{u}, n), (\mathbf{v}, n)) = \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_n}} \mu_{B_M}(\mathbf{u}, \mathbf{v}) = D_{\mathbf{u}}^{(I_n)}, \quad (26)$$

where the final  $D_{\mathbf{u}}^{(I_n)}$  is a combination coefficient for an  $I_n$ -truncation of a sum indexed from  $B_M$ . Clearly

$D_{(\mathbf{u},i)}^{(I)} = 0$  when  $i > n$ . When  $i < n$ ,

$$\begin{aligned} D_{(\mathbf{u},i)}^{(I)} &= \sum_{j=i}^n \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_j}} \mu_{B_M}(\mathbf{u}, \mathbf{v}) \mu_{[n]}(i, j) \\ &= \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_i}} \mu_{B_M}(\mathbf{u}, \mathbf{v}) - \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_{i+1}}} \mu_{B_M}(\mathbf{u}, \mathbf{v}) = D_{\mathbf{u}}^{(I_i)} - D_{\mathbf{u}}^{(I_{i+1})}, \end{aligned} \quad (27)$$

then (24) emerges by fixing  $n = 2$ , and choosing  $I_1 = \{\text{FCR}_{\text{LL}}\}$  and  $I_2 = \{\text{FCR}_{\text{HL}}\}$ . The coefficients  $p_{\mathbf{f}_i}^{\text{HL}}$  and  $p_{\mathbf{f}_i}^{\text{LL}}$  are recovered from (26) and (27), cf. [27, (17), (18)], and these also give a rederivation of (25).

Other choices of  $I_1$  and  $I_2$ , or, viewed alternatively,  $\{\text{FCR}_{\text{LL}}\}$  and  $\{\text{FCR}_{\text{HL}}\}$ , lead, after some straightforward manipulation and sometimes with recourse to the argument in Proposition 2.2, also to the working equations of a variety of other multilevel fragmentation methods. These include a multilevel version of the EE-MB [23], the HMBI [24], the MFBA [25], and the MC QM/QM method [26]. The full multilevel form of the simplex-based method in [20, 21] can also be recovered via the discussion of (21) above, and for more general  $n$ , the MIM [88] emerges. We omit the details for reasons of space, but see [47, Sec. 7.2]. Equivalences between some of these methods have certainly been mentioned before, as in, e.g., [24, 27], but this provides further support for the generality of the ML-FCR approach, at least for existing fragmentation methods.

Alternatively, given an interaction graph  $G$ , choosing in (23) instead  $\Pi = \text{conn}[G] \times \mathbb{N}$  delivers the ML-BOSSANOVA extension [14] to the original BOSSANOVA scheme. Very briefly, ML-BOSSANOVA expands each of a family of potentials  $\{V_p\}_{p \in \mathbb{N}}$  via a standard BOSSANOVA expansion as  $V_p = \sum_{\mathbf{u} \in \text{conn}[G]} \tilde{V}_{p,\mathbf{u}}$ . Each  $\tilde{V}_{p,\mathbf{u}}$  is itself decomposed as  $\tilde{V}_{p,\mathbf{u}} = \sum_{q=0}^p \tilde{\omega}_{q,\mathbf{u}}$ , with  $\tilde{\omega}_{q,\mathbf{u}} = \tilde{V}_{q,\mathbf{u}} - \tilde{V}_{q-1,\mathbf{u}}$  in general and  $\tilde{\omega}_{0,\mathbf{u}} = \tilde{V}_{0,\mathbf{u}}$  as a special case. The intention is to capture only the important terms from an ANOVA-like decomposition  $\sum_{\mathbf{u} \in B_M} \sum_{p \in \mathbb{N}} \tilde{V}'_{p,\mathbf{u}}$  where instead each  $\tilde{V}'_{p,\mathbf{u}}$  is defined instead for the full  $B_M$  rather than just  $\text{conn}[G]$ . But crucially, the issue of combination-consistency arises here again, just as for BOSSANOVA. It is easy to see that a direct product  $P \times Q$  is a meet semilattice if and only if both  $P$  and  $Q$  are, and so this particular choice of  $\Pi$  will not generally be combination-consistent with  $B_M \times \mathbb{N}$ . So, for most cyclic interaction graphs  $G$ , at least one possible  $I$ -truncation of the ML-BOSSANOVA expansion will not lead to a downward-closed truncation of the full ANOVA-like expansion  $\sum_{\mathbf{u} \in B_M} \sum_{p \in \mathbb{N}} \tilde{V}'_{p,\mathbf{u}}$ . Note that, correspondingly, the ML-FCR cannot always recover an arbitrary ML-BOSSANOVA  $I$ -truncation.

All of the multilevel methods discussed above consider only a single axis for the level of model quality. However, using (23), it is straightforward to extend this to arbitrarily many axes. The extension can be motivated as allowing the approximation of each  $\tilde{V}_{\mathbf{u}}$  in an MBE in the style of a composite method, following the combination-technique understanding in [6, 14]. The starting point is thus the Boolean algebra  $B_M$  corresponding to an ANOVA-like decomposition of  $V$ . The next axis,  $P_2 = [2N - 1]$ , indexes the various levels of *ab initio* theory outlined in (3) above. Then, like ML-BOSSANOVA, we introduce a third axis  $P_3 = \mathbb{N}$ , which indexes members of a systematically-improving family of basis sets. Again under appropriate assumptions regarding convergence, the resulting poset grid  $\Pi = B_M \times P_2 \times P_3$  produces an exact expansion of the Born-Oppenheimer potential as

$$V^{\text{BO}} = \sum_{\mathbf{p} \in \Pi} \tilde{V}_{\mathbf{p}} = \sum_{\mathbf{u} \in B_M} \sum_{m=1}^{2N-1} \sum_{p=1}^{\infty} \tilde{V}_{(\mathbf{u},m,p)}. \quad (28)$$

The first axis can of course be restricted to a subposet of  $B_M$ , or isomorphically, a subposet of induced subgraphs of some  $G$ . For combination-consistency to hold, this must be a meet subsemilattice of  $B_M$ , for example  $\mathcal{M}_g[G]$ . Taking the subgraph perspective, we refer to (28) as an *ML-SUPANOVA decomposition*.

### 3. Adaptive Algorithm

Standard  $n$ -body truncations of MBEs become for increasing  $n$  at best inefficient, and at worst numerically unstable [11, 28], and so choosing the correct value of  $n$  for use in practice is difficult. When considering

---

**Algorithm 1** Adaptive algorithm for calculation of order ideal  $I$  (adapted from [47, Alg. 3.3])

---

```

1:  $I, A \leftarrow \emptyset$ 
2:  $Q \leftarrow$  an empty priority queue
3:  $D, L, E \leftarrow$  empty (zero) tensors
4:  $i \leftarrow -1$ 
5: repeat
6:    $i \leftarrow i + 1$ 
7:   if  $i = 0$  then
8:      $I_{\text{new}} \leftarrow \{\hat{0}_\Pi\}$ 
9:   else
10:     $I_{\text{new}} \leftarrow$  new elements from entries of  $Q$  according to selection strategy
11:     $I \leftarrow I \cup I_{\text{new}}$ 
12:    for all  $\mathbf{p} \in I_{\text{new}}$  do
13:       $L_{\mathbf{p}} \leftarrow \mathcal{L}[V_{\mathbf{p}}]$ 
14:      Evaluate Möbius tensor  $M^{(\mathbf{p})}$ .
15:       $D \leftarrow D + M^{(\mathbf{p})}$   $\triangleright$  Update the full combination tensor,
16:       $E \leftarrow E + M^{(\mathbf{p})}$   $\triangleright$  Update the error-indicator tensor.
17:      Insert  $\mathbf{p}$  into  $Q$ , keyed by  $\|\mathcal{L}[\tilde{V}_{\mathbf{p}}](= \text{REDUCE}(M^{(\mathbf{p})} \odot L))\|/\mathcal{C}(\mathbf{p})$ .
18:     $\triangleright$  Update generating antichain, and correct the error indicator tensor  $\triangleleft$ 
19:     $A \leftarrow A \cup I_{\text{new}}$ 
20:     $R \leftarrow A \cap \bigcup_{\mathbf{p} \in I_{\text{new}}} \{\mathbf{q} \in \Pi \mid \mathbf{q} \prec \mathbf{p}\}$ 
21:     $A \leftarrow A - R$ 
22:     $E \leftarrow E - \sum_{\mathbf{p} \in R} M^{(\mathbf{p})}$ 
23:     $\triangleright$  Calculate approximation, error indicator, cost, and uncertainty  $\triangleleft$ 
24:     $S_i \leftarrow \text{REDUCE}(D \odot L)$ 
25:     $\mathcal{E}_i \leftarrow \text{REDUCE}(E \odot L)$ 
26:     $\mathcal{C}_i \leftarrow \sum_{\mathbf{p} \in I} \mathcal{C}(\mathbf{p})$ 
27:     $\text{d}S_i \leftarrow \sqrt{\text{REDUCE}(\varepsilon^2(D \odot D))}$ 
28:  until  $Q$  is empty, or termination criteria are met.
29: return  $I$ , and also  $S_i, \mathcal{E}_i, \mathcal{C}_i$ , and/or  $\text{d}S_i$  as required

```

---

expansions like (16) and especially multilevel forms like (28), the task of choosing suitable truncations *a priori* becomes even harder again. Algorithm 1 outlines instead an adaptive, *a posteriori* approach for the selection of a truncation order ideal  $I$  for a sum (23) over some poset grid  $\Pi$ . This algorithm is a specialization of one given in a more general setting in [47, Chap. 3]. The basic idea of the algorithm goes back to work in dimensionally-adaptive numerical quadrature [38], and there are many other strong connections in the context of sparse grid approximation and the combination technique; see, e.g., [37, 39, 40]. We observe with interest that a similar approach is taken in the “bottom-up” energy-screening algorithm described very recently in [89], which walks a directed acyclic graph constructed for the terms of the GMBE. As there noted, this graph allows interpretation as a Hasse diagram, and Algorithm 1 can likewise be interpreted as walking the Hasse diagram of  $\Pi$ , in the more general order-theoretic sense [32]; see [47, Sec. 3.5], and cf., e.g., [40]. Our algorithm targets more general expansions than just the GMBE, and differs particularly in the screening criterion applied and the update mechanism; cf. [89, (12), (13)].

We assume here the choice of a linear operator  $\mathcal{L}$ , e.g., a point evaluation operator as in Section 2.1. Sequences of order ideals  $I^{(i)}$  and their corresponding evaluated sums  $S_i := \mathcal{L}[S_{I^{(i)}}]$  are generated, starting from the initial values  $I^{(0)} = \{\hat{0}\}$  and  $S_0 = \mathcal{L}[\tilde{V}_0] = \mathcal{L}[V_0]$ . At the  $i$ th iteration of the algorithm, at least one element  $\mathbf{p} \in I^{(i-1)}$  is selected for *expansion*: all those elements  $\mathbf{q} \succ \mathbf{p}$  such that  $I^{(i)} \cup \{\mathbf{q}\}$  would remain an order ideal are adjoined to  $I^{(i-1)}$  to form  $I^{(i)}$ . The contribution potentials  $\tilde{V}_{\mathbf{q}}$  for these new element(s) are evaluated, and accumulated into  $S_i$ .

The choice of element(s)  $\mathbf{p} \in I^{(i-1)}$  for expansion is guided by the benefit/cost ratios  $\|\mathcal{L}[\tilde{V}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$ , where  $\mathcal{C} : \Pi \rightarrow \mathbb{R}$  is an *abstract cost model* for the evaluation of the subproblem potentials  $V_{\mathbf{p}}$  by  $\mathcal{L}$ . One possible abstract cost model for point evaluation is outlined in Appendix B. Under certain assumptions, see [14, 37] and [47, Sec. 3.4], a truncation constructed by assembling terms in descending order of

benefit/cost leads to a quasi-optimal truncation of (23).

Notationally, we assume that the elements of a  $d$ -dimensional tensor, i.e., a multidimensional array  $T$  can be indexed  $T_{\mathbf{m}}$  by multiindices  $\mathbf{m} \in \mathbb{N}^d$ . For a poset grid  $\Pi = P_1 \times \cdots \times P_d$  where each  $P_i$  is countable, we assume the existence of a suitable indexing bijection  $\Phi : \Pi \rightarrow \mathbb{N}^d$  and will write by abuse of notation  $T_{\mathbf{p}}$  to mean the element  $T_{\Phi(\mathbf{p})}$ . Elementwise additions (+) and multiplications ( $\odot$ ) of tensors behave as usual, and the tensor product  $T \otimes U$  of a  $k$ -dimensional  $T$  and  $l$ -dimensional  $U$  is  $(k+l)$ -dimensional and defined elementwise by  $(T \otimes U)_{(m_1, \dots, m_k, n_1, \dots, n_l)} := T_{(m_1, \dots, m_k)} U_{(n_1, \dots, n_l)}$ . We write  $\text{REDUCE}(T) := \sum_{\mathbf{m} \in \Pi} T_{\mathbf{m}}$ . This sum is here always finite, since although tensors may be notionally infinite, those considered have finitely many non-zero terms by construction. For the same reason, it is desirable to work in the implementation with sparse tensors [90, 91], that is, higher-dimensional analogues of standard sparse matrices. We omit the details here.

It is a deliberate feature of this notation that the combination coefficients  $D_{\mathbf{p}}^{(I^{(i)})}$  as per (14) coincide identically with the entries of the *combination tensor*

$$D^{(I^{(i)})} := \sum_{\mathbf{p} \in I^{(i)}} M^{(\mathbf{p})},$$

where each  $M^{(\mathbf{p})}$  is the *Möbius tensor* of  $\mathbf{p} \in \Pi$  and is defined elementwise by  $M_{\mathbf{q}}^{(\mathbf{p})} := \mu_{\Pi}(\mathbf{q}, \mathbf{p})$ . As elements  $\mathbf{p} \in \Pi$  are added to the order ideals  $I^{(i)}$ , their evaluations  $\mathcal{L}[V_{\mathbf{p}}]$  are stored in a suitable tensor as  $L_{\mathbf{p}}$ . Their Möbius tensors can be used to directly evaluate  $\mathcal{L}[\tilde{V}_{\mathbf{p}}] = \text{REDUCE}(M^{(\mathbf{p})} \odot L)$ , and then accumulated into a running combination tensor  $D$ .

Maintaining the sum  $S_i$  indirectly via the combination tensor has some appealing numerical characteristics that help to ameliorate issues like those noted in, e.g., [28]. A sparsity-exploiting evaluation  $S_i = \text{REDUCE}(D \odot L)$  can use as few arithmetic operations as possible, whereas a naïve accumulation of evaluated terms  $\mathcal{L}[\tilde{V}_{\mathbf{p}}]$  may suffer over iterations from accrued error due to floating-point arithmetic. Moreover, all *ab initio* methods reduce to the solution of particular systems of equations, which are solved to within some tolerance usually chosen well below floating-point precision. In [28], a propagation-of-errors technique is used to test how these tolerances flow through to the overall uncertainty of a standard  $n$ -body truncation, rewritten for the purpose in a closed-form expression [28, (4.2) and (4.5)]. If we associate to each  $\mathcal{L}[V_{\mathbf{p}}]$  an inherent uncertainty  $\varepsilon_{\mathbf{p}}$ , then it is an easy generalization of the same approach to calculate the propagated uncertainty in the sum  $S_I$  for any  $I$  as

$$\text{d}S_I = \sqrt{\sum_{\mathbf{p} \in I} \left( D_{\mathbf{p}}^{(I)} \right)^2 \varepsilon_{\mathbf{p}}^2}.$$

The precise encoding of poset grid elements  $\mathbf{p} \in \Pi$  is an implementation detail. The algorithm requires only that all covered elements  $\mathbf{q} \prec \mathbf{p}$  can be enumerated, and similarly all covering elements  $\mathbf{r} \succ \mathbf{p}$ . It suffices if the same is possible for any element of an individual axis  $p \in P_i$ . This is easy in the cases of  $[n]$ ,  $\mathbb{N}$ , and  $B_n$ . A little more work is required for  $\text{conn}[G]$  and particularly  $\mathcal{M}_g[G]$ ; see [47, App. 2] for details and supporting references. The values  $\mu_{\Pi}(\mathbf{q}, \mathbf{p})$  and thus  $M^{(\mathbf{p})}$  can then always be evaluated explicitly at least as per (10), which is amenable to memoization. Particularly when expressions like (11) and (12) are available, it is more convenient to apply (22) and write

$$M^{(\mathbf{p}=(p_1, \dots, p_d))} = \bigotimes_{i=1}^d m^{(p_i)},$$

using Möbius “vectors” defined elementwise by  $m_s^{(p_i)} := \mu_{P_i}(s, p_i)$  for  $s, p_i \in P_i$ . For example, if  $P_i$  is a chain poset, then constructing  $m^{(p_i)}$  requires no more than two elementwise updates to an initially-zero sparse vector.

We maintain throughout the course of the algorithm a priority queue, see [78] and cf. [38], of all elements  $\mathbf{p} \in \Pi$  which have been previously added to  $I$  and not yet fully expanded, in the sense that not all elements  $\mathbf{q} \succ \mathbf{p}$  have also been added to  $I$ . The queue is kept in descending order of  $\|\mathcal{L}[\tilde{V}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$ . At line 10, elements  $\mathbf{p}$  are removed one-by-one from the front of the queue, and their covering elements  $\mathbf{r} \succ \mathbf{p}$  are tested for *admissibility*, which holds whenever all elements  $\mathbf{p}' \prec \mathbf{r}$  are also already in  $I$ . Any such

admissible  $\mathbf{r}$  joins the overall set of new elements  $I_{\text{new}}$ , such that  $I^{(i)} \leftarrow I^{(i-1)} \cup I_{\text{new}}$ . A  $\mathbf{p}$  with at least one admissible  $\mathbf{r} \succ \mathbf{p}$  is *expandable*, and different strategies can be applied to guide the number of elements expanded at each step. One might simply stop after expanding the first, BEST expandable element in the queue, or instead consider ALL expandable elements in the queue. A more flexible strategy is to continue through the queue, expanding any elements whose benefit/cost ratio is within some THRESHOLD factor  $0 \leq \alpha \leq 1$  of that of the best. Whichever strategy is used, any dequeued element found to have an inadmissible  $\mathbf{r} \succ \mathbf{p}$  should be requeued.

Once the new elements  $\mathbf{p} \in I_{\text{new}}$  have been identified, each corresponding  $\mathcal{L}[\mathbf{V}_{\mathbf{p}}]$ ,  $M^{(\mathbf{p})}$ ,  $\mathcal{L}[\tilde{\mathbf{V}}_{\mathbf{p}}]$ , and  $\mathcal{C}(\mathbf{p})$  must be evaluated, the combination tensor  $D_{\mathbf{p}}^{(I)}$  updated, and  $\mathbf{p}$  inserted into the queue. This occurs in the loop at line 12, along with one additional step which will be explained below. Parallelism in this loop should be exploited as much as possible, since in practice, the evaluations  $\mathcal{L}[\mathbf{V}_{\mathbf{p}}]$  are by far the most costly part of the algorithm.

The algorithm terminates once either the queue has been exhausted, or some termination criterion has been met. One might place a cumulative limit on the abstract costs  $\mathcal{C}(\mathbf{p})$ , but it is more useful to have an estimating *error indicator*, see and cf., e.g., [37, 38]. For this, we suggest

$$\mathcal{E}_i = \sum_{\mathbf{p} \in A^{(i)}} \mathcal{L}[\tilde{\mathbf{V}}_{\mathbf{p}}], \quad (29)$$

where  $A^{(i)}$  is the set of maximal elements of  $I^{(i)}$ . For efficiency, the adaptive algorithm also maintains and updates this set. The key observation is that any new  $\mathbf{p}$  added to  $I^{(i)}$  must be a maximal element of the same, and that any elements  $\{\mathbf{q} \in I^{(i-1)} \mid \mathbf{q} \prec \mathbf{p}\}$  are correspondingly precluded from maximality in  $I^{(i)}$ . Again, we calculate (29) not as a sum of contribution terms, but rather maintain a set of combination coefficients analogous to those for the full sum (lines 16 and 19–22).

## 4. Experiments and Discussion

As an initial evaluation of the techniques outlined above, we consider a collection of experimental results obtained by adaptively growing truncated sums of the form (23). These results are a subset of those covered in [47], and were generated using an implementation of the more general version of Algorithm 1 given there.

A detailed discussion of implementation details is not necessary to interpret these results, and is left to [47]. Similarly out of scope is a discussion of the computational treatments of the various involved posets; for this, see particularly Appendix B of that source. Briefly, however, the implementation made use of standard scientific libraries [92, 93], particularly NetworkX [94] for working with graphs. Some reference was made to the Sparse library for sparse arrays [91]. In general, tensor reductions were performed using interval arithmetic as implemented in the Arb library [95], using 100 bits of precision, but we believe this had no significant impact on results compared to the use of standard double-precision floating-point arithmetic. Some of the data in Figure 5 were regenerated to correct an error in [47], and here, standard double-precision was used. Plots and visualizations are colored using the “bright” scheme of [96], while molecular visualizations were rendered in Blender using the DSO shading scheme in [97].

Results given here were obtained using HF [1], MP2 [51], CCSD [98, 99], and CCSD(T) [54] calculations; KS-DFT/B3LYP [100] was used for geometry optimizations. We used variously the following basis sets: STO-3G [101, 102], cc-pVDZ [58, 103], cc-pVTZ, cc-pCVnZ for  $2 \leq n \leq 8$  [58, 104–114], and also all necessary basis sets for G4(MP2) [60, 61, 115–118]. Most basis sets were obtained from the Basis Set Exchange [119], and processed for computational efficiency using the BSE Python library. For particularly fine details on basis sets and, e.g., detailed solver settings, see again [47].

Calculations required for subproblem potential evaluation were performed using PySCF [120–122], to an iterative convergence threshold set at  $10^{-8} E_h$ , with a suitably tight integral prescreening threshold. Full-system geometry optimizations, as well as one reference calculation, were performed using NWChem [123]. Monatomic energies used to calculate atomization energies, see below, were mostly obtained using MRCC [124–128], with HF and CC convergence thresholds at  $10^{-10} E_h$ ; those for oxygen with aug-cc-pCV8Z were obtained using PySCF, converged to  $10^{-9} E_h$ . An implicit assumption here is that single-point results for identical geometries and levels of theory but carefully obtained under different solver packages will be reliably equivalent up to convergence thresholds.

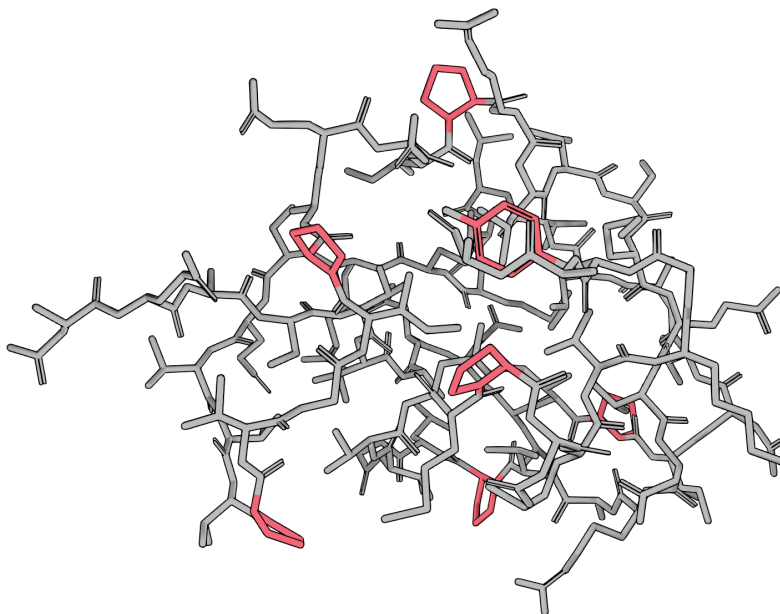


Figure 1: Stick-model visualization of the covalent bond structure of the antifreeze protein 1KDF [129, 130], after preprocessing as described in the main text and with hydrogen atoms excluded. Ring substructures are highlighted in red.

In the following, we are primarily interested in observing changes in the accuracy of the sums at each iteration against their total abstract costs of calculation  $C_i$ , as per line 26 of Algorithm 1 and the abstract cost model outlined in Appendix B. With one exception, discussed below, the final plots given here were generated by reusing previously-cached subproblem potential evaluations. Since these plots consider only abstract cost, this reuse has no impact on the results as shown.

#### 4.1. Convex SUPANOVA decomposition

Before considering the full multilevel setup, we first investigate the suitability of the convex SUPANOVA correction to the non-multilevel BOSSANOVA method. Since we already know from, e.g., [18, 19] that the conventional BOSSANOVA approach is well capable of handling smaller chain-like molecules, we investigate two examples of larger molecular systems with characteristics that would be problematic in light of BOSSANOVA’s issues with combination-consistency.

We begin with a model of an *antifreeze protein* [129], obtained from the Protein Data Bank [130, PDB key: 1KDF]. As for most proteins, and as indicated in the visualization in Figure 1, the covalent bond structure of 1KDF contains several rings. The 1KDF model was preprocessed using Open Babel [131], with hydrogens added explicitly. The resulting system contained a total of 991 atoms, 479 non-hydrogen. Using NWChem, via a calculation initialized by basis set projection from an initial HF/STO-3G calculation and then converged to  $10^{-9} E_h$ , we obtained a reference HF/cc-pVTZ total energy of the 1KDF system as approximately  $-24\,896.350 E_h$ .

That some interaction graphs may not accurately represent the spatial topology of a molecule is a known issue in graph-based fragmentation methods [22]. The shape of 1KDF suggests that strong interactions should be expected between collections of particles that are close in space, but not close in the standard bond graph  $G$ . The impacts of those interactions would then only appear in contribution terms  $\tilde{V}_{\mathbf{u}}$  corresponding to quite large convex subgraphs  $G[\mathbf{u}]$  which also include at least all those particles along a shortest path between the two originally considered. As mitigation, we employed instead an interaction graph  $G$  with a distance-thresholded edge set  $E = \{\{i, j\} \subseteq [M] \mid i \neq j, \|R_i - R_j\| \leq r_{\text{cut}}\}$ . This is a common approach in fragmentation methods, either explicitly or implicitly; see and cf., e.g., [20–22], as



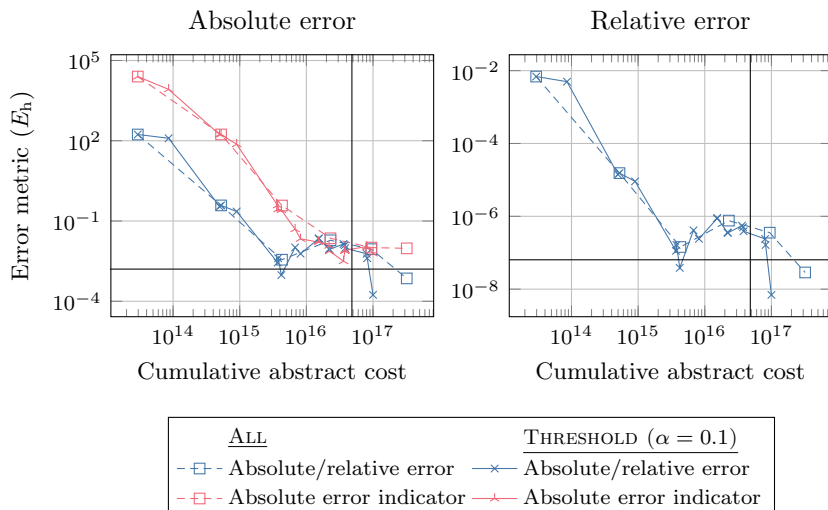


Figure 2: Per-iteration error metrics for adaptive truncations of a convex SUPANOVA decomposition for the 1KDF antifreeze protein, obtained using ALL and THRESHOLD selection strategies. The horizontal black line in each plot indicates chemical accuracy,  $1 \text{ kcal mol}^{-1}$ . The vertical black line indicates the abstract cost of a full-system reference HF/cc-pVTZ calculation.

well as many other examples in the reviews cited above. Here, we used a cutoff radius of  $r_{\text{cut}} = 2.5 \text{ \AA}$ , which should be sufficient to capture at least any hydrogen bonds [132] in addition to all covalent bonds.

In the convex SUPANOVA decomposition (16), we use notional subproblem potentials  $V_{\mathbf{u}}$  such that an evaluated  $\mathcal{L}[V_{\mathbf{u}}]$  is the HF/cc-pVTZ total energy of the subsystem indexed by  $\mathbf{u}$ , calculated using PySCF. Dangling covalent bonds from subsystems were treated with standard hydrogen link atoms, placed using covalent radii [133] consistent with, e.g., [13, (19)]. Rather than a nuclear decomposition, we used a fragment decomposition. The choice of fragments was intended to avoid two practical issues: firstly, the question of how to terminate a dangling double bond, and second, the possibility that two link atoms would be placed in close spatial proximity after subsystem excision, thus potentially biasing the relevant  $\tilde{V}_{\mathbf{u}}$ ; on the latter, cf., e.g., [22, 75]. To this end, we refined an initial fragmentation  $F = \{\{i\}\}_{i=1}^M$  of  $[M]$  using a simple heuristic algorithm, the details of which are given in Appendix C. Technically, this means that the decomposition is not in terms of  $\mathcal{M}_g[G]$ , but rather  $\mathcal{M}_g[G']$ , where  $G' = G/F$  is the *quotient graph* [134]  $G' = (F, E')$  with edge set  $E' = \{\{F_{i'}, F_{j'}\} \mid i' \neq j', \exists \{i, j\} \in E \text{ with } i \in F_{i'}, j \in F_{j'}\}$  derived from the original edge set  $E$  of  $G$ .

Per-iteration results are plotted in Figure 2 for two adaptively-grown truncations of the convex SUPANOVA decomposition. One truncation was refined using the ALL selection strategy, and the other according to the THRESHOLD strategy with  $\alpha = 0.1$ . Both calculations were terminated once the total, cumulative abstract cost expended passed the abstract cost of the reference calculation by a factor of two. The left-hand plot shows both the progression of the error indicators  $\mathcal{E}_i$  at each iteration, as well as the true absolute error of the resulting value  $S_i$  as measured against the reference total energy. The right-hand plot shows the per-iteration relative errors for the same calculations. The solid black vertical and horizontal lines on the plots indicate the abstract cost of the reference calculation, and the  $1 \text{ kcal mol}^{-1}$  threshold of chemical accuracy relative to reference, respectively. Since the first iteration of the algorithm produces in both cases only  $S_0 = \mathcal{L}[V_{\emptyset}] = 0$  for zero abstract cost, these data points are not explicitly shown, nor will they be in similar plots to follow.

The absolute errors of the produced sums  $S_i$  under both ALL and THRESHOLD selection strategies decrease rapidly in early iterations, stabilizing around  $10^{-2} E_h$ . A further improvement is suggested shortly after the abstract costs of the iterative calculations pass that of the reference calculation. Excepting one early THRESHOLD iteration, chemical accuracy relative to the reference value is only achieved in the final iteration. From a performance perspective, we would ideally hope to achieve chemical accuracy at a significantly reduced cost relative to the reference calculation. Although we do not achieve this, it should

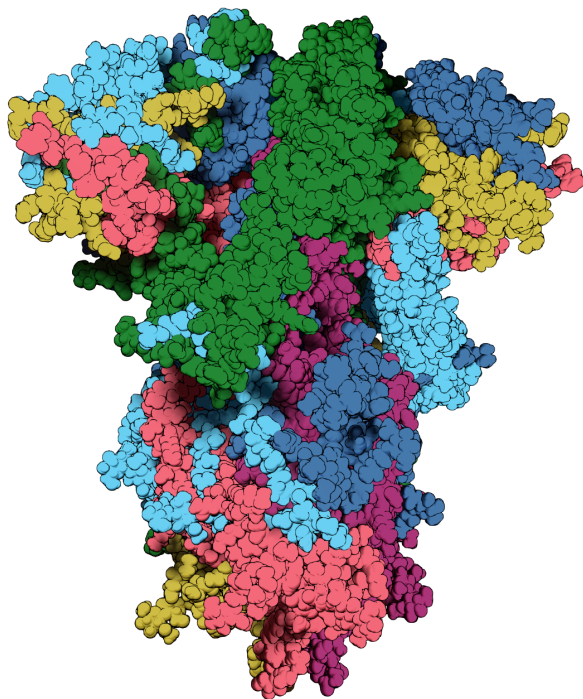


Figure 3: Space-filling visualization of the SARS-CoV-2 spike glycoprotein [135, 136, PDB key: 6VXX]. Atoms are drawn as spheres of appropriate van der Waals radii [137]. Colors indicate membership of connected components of the full-system bond graph, chosen using a graph-coloring algorithm provided by NetworkX.

be kept in mind that, as plot (b) highlights, a consistent relative accuracy of at least six significant figures is achieved with approximately an order of magnitude speedup relative to reference. For larger systems again, such reliable relative accuracy is probably a more important consideration than true chemical accuracy, especially for total energy calculations.

As expected, there is no indication here of the inherent errors that would be expected in higher-order truncations of the standard BOSSANOVA expansion due to the presence of chordless cycles in  $G'$ . It is also positive to see that the error indicators  $\mathcal{E}_i$  are generally reliable. If anything, they are a little conservative: the true error is sometimes significantly overestimated, but only rarely underestimated, and even then not by more than an order of magnitude. Further, although not plotted explicitly in Figure 2, the propagated uncertainties  $dS_i$  in these calculations hovered generally around  $10^{-6} E_h$ , and never exceeded  $10^{-5} E_h$ , assuming throughout a per-evaluation uncertainty of  $\varepsilon = 10^{-8} E_h$  consistent with involved convergence thresholds. This suggests that the unfavorable error propagation observed in [28] for  $n$ -body truncations of the MBE is not in play here. It seems unlikely that this is due to any particular combinatorial property of the decomposition itself, but rather only that most terms  $V_{\mathbf{u}}$  in the underlying MBE have zero combination coefficients; this would be consistent with, e.g., [28, 67].

As a general feasibility test of the convex SUPANOVA approach, we performed a similar calculation on a much larger sample system, specifically, a model of the spike glycoprotein of the SARS-CoV-2 virus [135], again drawn from the Protein Data Bank [136, PDB key: 6VXX]. Preprocessing and hydrogenation with Open Babel produced here a system made up of 27 non-covalently bonded subsystems, each containing between 63 and 2690 non-hydrogen atoms, for a total of 46 923 atoms. A space-filling visualization of the 6VXX system is given in Figure 3. We used the same methodology as above to choose a nuclear interaction graph  $G$  ( $r_{\text{cut}} = 2.5 \text{ \AA}$ ), a refined fragmentation  $F$ , and a final quotient graph  $G' = G/F$ . This  $F$  included 7524 fragments, each containing up to 17 distinct atoms.

We performed a single adaptive calculation, using HF/cc-pVTZ subproblem potentials as above, and

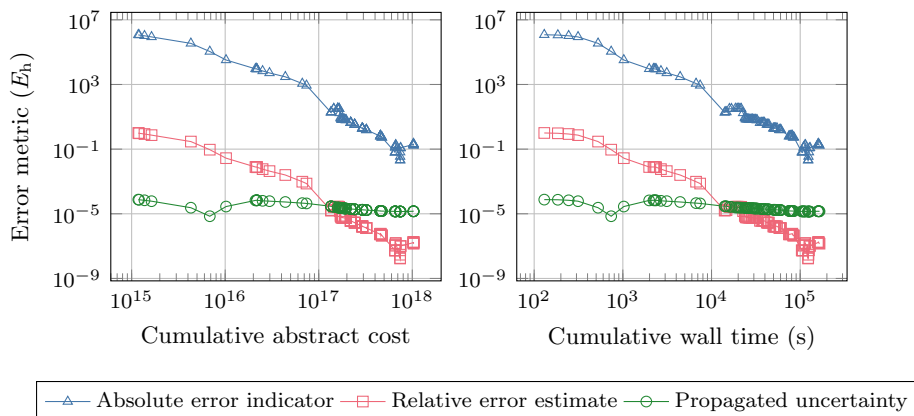


Figure 4: Per-iteration error metrics for adaptive convex SUPANOVA approximation of the HF/cc-pVTZ total energy of the SARS-CoV-2 spike glycoprotein (PDB key: 6VXX) [135, 136].

applying the THRESHOLD strategy with  $\alpha = 0.5$ . The 6VXX system is of course far too large to allow a conventional reference calculation at any level of quantum theory. Indeed, we were not even able to evaluate an abstract cost for the complete system, since our abstract cost model involves the explicit calculation of some  $\mathcal{O}(N_{\text{AO}}^2)$  two-electron integrals; see Appendix B. Thus, we set no explicit termination criteria, and simply allowed the calculation to run for approximately 46 h. At each iteration, subproblem potential evaluations were distributed in a hybrid fashion across up to 4864 cores of a compute cluster.

Results for this calculation are shown in Figure 4. In the absence of a reference result, we can only consider the per-iteration error indicators. The left plot shows the absolute values of these as a function of cumulative abstract cost, also an estimate of the relative error calculated as  $|\mathcal{E}_i/S_i|$ , and the propagated uncertainty of the calculation assuming per-subproblem thresholds of  $\varepsilon = 10^{-8} E_h$ . The right-hand pane plots the same data, but measured as a function of the true cumulative wall-clock time required for the calculation. The similarity of the two plots supports the “reasonableness” of the abstract cost model we have used throughout, at least for HF. Some deviance here can be attributed to unavoidable parallel inefficiency.

The error indicators exhibit a clear and fairly smooth rate of decay. At termination, they suggest the calculation has approximated the total energy of the spike glycoprotein to again at least six significant figures of accuracy. Naturally, this conclusion depends entirely on the reliability of the error indicator. Although this is certainly suggested by results in the smaller 1KDF case, we do mention that we discussed cases in [47] where the error indicator underperformed for reasons related to unrepresentative interaction graphs, and so further and more detailed validation would be in order before this could be relied upon. What can be said, however, is that the propagated uncertainty of this calculation remains basically flat throughout, and in particular, always stay several orders of magnitude less than the absolute value of the error indicator. So the numerics of the truncated sum are sound.

This test provides initial support to the idea that the convex SUPANOVA approach can be practically applied to extremely large molecules. It is to be stressed, however, that this is a preliminary calculation only, and further benchmarking of the underlying model is most certainly required. One potential target for such benchmarking is provided by an FMO calculation performed on the same protein [138]. That calculation was much more methodologically sophisticated, and included in particular a high-quality treatment of correlation energy, so a direct comparison is not appropriate without further work. We do note that the computational resources applied in that study were substantially greater than ours, and it would be very interesting to investigate whether any computational advantage remained once all differentiating factors were properly taken into account.

## 4.2. ML-SUPANOVA decompositions

When evaluating experimental results obtained by truncating a decomposition like (28) of the true Born-Oppenheimer potential  $V^{\text{BO}}$ , we are impeded by the fact that there is no non-trivial molecular

system for which a true reference solution can be obtained. We consider here only much smaller test systems than in the previous section. For these, we can obtain reasonable-quality approximate solutions, and these can function at least as rough references. Specifically, we investigate calculations on the linear alkane heptane ( $\text{C}_7\text{H}_{16}$ ), and the heterocyclic molecule limonin ( $\text{C}_{26}\text{H}_{30}\text{O}_8$ ). Chemical structures for these systems were obtained from the ChemSpider database [139, 140], and then relaxed into plausible equilibrium conformations using a KS-DFT/B3LYP/cc-pVDZ geometry optimization under NWChem, using the SIMINT ERI calculation library [141].

The poset grid under theoretical consideration is  $\Pi = \mathcal{M}_g[G] \times [2N - 1] \times \mathbb{N}$ , as outlined above, but see the practical restriction discussed below. It is also possible to construct decompositions in terms of “subgrids” of this poset grid. These can be viewed as holding one or multiple of the parameters  $\mathbf{u}$ ,  $m$ , and/or  $p$  for potentials  $V_{(\mathbf{u},m,p)}$  fixed, and doing so allows us to investigate the utility of the various axes. We consider the following subgrids of  $\Pi$ :

- $\Pi_{\text{non-ML}} = \mathcal{M}_g[G]$ , with subproblem potentials fixed at a particular level of theory indexed by  $(m, p)$ . This is a non-multilevel approximation. In cases where  $\mathcal{M}_g[G] = \text{conn}[G]$ , this is a BOSSANOVA decomposition; where not, a more general convex SUPANOVA one.
- $\Pi_{\text{GCM}} = [2N - 1] \times \mathbb{N}$ . All subproblem potentials are full-system potentials  $V_{m,p}$ , that is, with  $\mathbf{u} = [M]$ . Consistent with [47], where the idea was considered by extension of [6], we refer to this as a GCM (*generalized composite method*) grid.
- $\Pi_{\text{AbI}} = \mathcal{M}_g[G] \times \mathbb{N}$ , with subproblem potentials fixed at a particular *ab initio* method indexed by  $m$ , e.g.,  $m = 4$  for CCSD(T). As above, for cases where  $\mathcal{M}_g[G] = \text{conn}[G]$ , this is an ML-BOSSANOVA decomposition.
- $\Pi_{\text{BS}} = \mathcal{M}_g[G] \times [2N - 1]$ , with subproblem potentials using a fixed basis set indexed by  $p$ , e.g.,  $p = 2$  for cc-pCVTZ.

For each test case, we calculated a reference total energy using a single-point calculation via PySCF, at as high a level of theory as practically possible. We also used an alternative mechanism to derive a reference result, specifically, an implementation of the G4(MP2) method [61], written in Python with some reference to [142]. We omitted here the initial DFT-based geometry optimization technically required by G4(MP2). In general, the component calculations required were also performed using PySCF, using monatomic energies generally calculated using MRCC, as described above.

Since G4(MP2) makes the frozen-core approximation, it is inherently unsuited to the calculation of accurate total energies, but might be reliable to around chemical accuracy [143, 144] for the calculation of energy differences such as atomization energies. An atomization energy potential can be constructed from the Born-Oppenheimer potential as

$$V^{\text{atom}}(X_1, \dots, X_M) := \left[ \sum_{A=1}^M V^{\text{BO}}(X_A) \right] - V^{\text{BO}}(X_1, \dots, X_M), \quad (30)$$

cf., e.g., [4]. Similar potentials can also be derived from approximations to the Born-Oppenheimer potential like any of the  $V_{(\mathbf{u},m,p)}$  terms implicitly involved in (28), taking care to match the level of theory  $(m, p)$  for the monatomic calculations. Since monatomic energies  $V_{m,p}(X_A)$  can be precalculated and reused, we assume that the abstract costs of such potentials are precisely the same as the standard total-energy versions.

For practical reasons, we restricted the poset axes  $[2N - 1]$  and  $\mathbb{N}$  for *ab initio* theory and basis set. For atomization energy calculations, we restricted the *ab initio* theory axis to contain four elements corresponding to HF, MP2, CCSD, and CCSD(T), and the basis set axis to contain seven elements, corresponding to cc-pCV $n$ Z for  $n = 2$  (cc-pCVDZ) through 8 (cc-pCV8Z). For total energy calculations, we further restricted these axes according to the level of theory used to derive the reference result. While we would anticipate larger axes to produce more accurate approximations of  $V^{\text{BO}}$ , any additional accuracy would confound a comparison against the reference result. For atomization energy calculations, we do not consider the subgrids  $\Pi_{\text{non-ML}}$  or  $\Pi_{\text{BS}}$ ;  $\Pi_{\text{AbI}}$  is, however, considered, fixed for CCSD(T).

The reference all-electron CCSD(T)/cc-pCV5Z total energy of heptane was approximately  $-276.370 E_h$ , and the atomization energy of heptane as per G4(MP2) was approximately  $3.481 E_h$ . We used here only

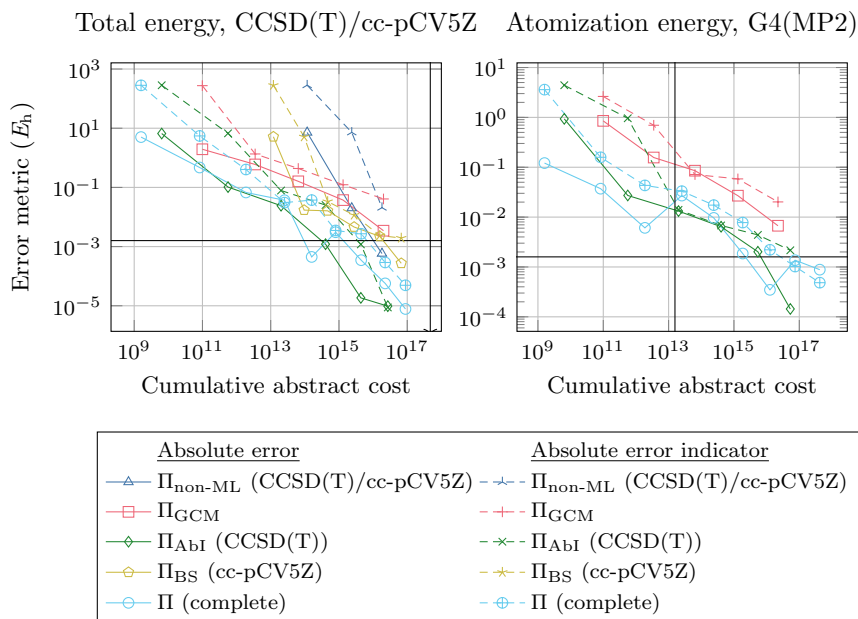


Figure 5: Per-iteration error metrics for adaptive truncations of the ML-SUPANOVA decomposition for heptane, and for subgrid decompositions.

the standard covalent bond graph  $G'$ . The corresponding fragmentation  $F$  of the nuclear indices, with each fragment  $F_i$  containing the indices of exactly one backbone carbon atom and also those of any hydrogen atoms bonded to it, is both as delivered by the heuristic algorithm, and precisely consistent with that suggested for the original BOSSANOVA decomposition. We applied here only the ALL selection strategy.<sup>3</sup> Plots of per-iteration error metrics for adaptive truncations of the complete poset grid  $\Pi$  and of the variously-considered subgrids listed above are given in Figure 5. The left-hand plot shows accuracy results for total energy calculations, as measured against the CCSD(T)/cc-pCV5Z reference, and corresponding error indicators. The right-hand plot is equivalent, but for atomization energy calculations, measured against the G4(MP2) reference result. We correct in this plot an error in [47, Fig. 7.1].<sup>4</sup>

For the total energy results, we see, in short, that all considered sequences of truncations lead (with the narrow exception of the GCM result) to chemically-accurate approximations of the reference result, and at speedups of at least an order of magnitude. It is interesting to observe that the generally best performance is delivered by the subgrid fixed at CCSD(T); since  $G$  is a simple chain, here  $\mathcal{M}_g[G] = \text{conn}[G]$ , and this is precisely an ML-BOSSANOVA expansion. The performance here is any case comparable with truncations for the complete poset grid  $\Pi$ , while the fixed-basis  $\Pi_{\text{BS}}$  grid performs less well. The error indicators for all calculations are generally well-behaved, although sometimes prone to overestimation of the true error, particularly for the  $\Pi_{\text{non-ML}}$  and  $\Pi_{\text{AbI}}$  cases.

For the calculation of atomization energies, no cost benefit is obtained relative to the reference calculation. We do not read much into this: heptane is still only a small system, and it is unsurprising that the CCSD and CCSD(T) calculations involved in the ML-SUPANOVA truncation are more expensive than the up-to-MP2 calculations required for G4(MP2). Again, the per-iteration absolute errors for the three tested grids eventually trend progressively downwards, although the results for the complete grid  $\Pi$  seem to spend considerable time in a pre-asymptotic regime. Truncations for both  $\Pi$  and  $\Pi_{\text{AbI}}$  consistently outperform those for  $\Pi_{\text{GCM}}$  at similar cost, and both come to approximate the G4(MP2) atomization energy to within chemical accuracy. A slight advantage is noticeable here for the complete

<sup>3</sup>Note that for the THRESHOLD and BEST strategies, some minor adjustment of Algorithm 1 would be required to handle the fact that, e.g.,  $\tilde{V}_{(\emptyset, m, p)} = 0$ .

<sup>4</sup>In [47], the absolute error values plotted were measured relative, not to the G4(MP2) atomization energy of heptane as stated, but rather the ccCA-PS3 [62] atomization energy of heptane, see [47, Tab. 7.2]. The error behaviour shown in the original [47, Fig. 7.1] seems, in this light, consistent with the reasoning that led to the choice of the G4(MP2) value as reference, rather than the ccCA-PS3 atomization energy.

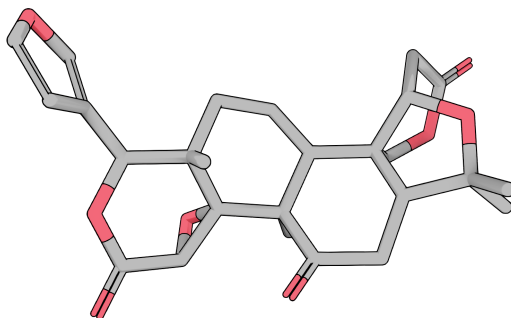


Figure 6: Stick-model visualization of limonin ( $C_{26}H_{30}O_8$ ), after KS-DFT/B3LYP/cc-pVDZ geometry optimization. Sticks indicate single and double covalent bonds between non-hydrogen atoms; hydrogen atoms themselves, and bonds to them, are not shown.

grid  $\Pi$ , and the final iterations for that grid suggest a slightly faster rate of “convergence” relative to those for  $\Pi_{AbI}$ . The absolute error indicators for all grids are generally reliable, tracking the true absolute error closely. When considering the absolute error and the absolute error indicator for  $\Pi$  in the final iterations, we should remember that the G4(MP2) result may itself not be reliable much past the chemical accuracy cutoff. It remains therefore possible that the error indicator is valid and the true FCI/CBS atomization energy of heptane is indeed being approximated more and more closely. This cannot be the case for  $\Pi_{AbI}$ , which can at best approximate the CCSD(T)/CBS atomization energy.

The second system we consider, limonin, is of interest particularly due to the highly cyclic nature of its covalent bond graph  $G'$ , clearly visible in the visualization in Figure 6. Cyclic structures remain after fragmentation with the heuristic algorithm: the resulting quotient graph  $G' = G/F$  contains five chordless cycles of length at least five. This indicates that the standard ML-BOSSANOVA decomposition using  $\text{conn}[G]$  cannot succeed, and the use of  $\mathcal{M}_g[G]$  is more appropriate; although we do not give details here for reasons of space, this was established for the single-level BOSSANOVA case in [47, Sec. 6.6]. It was infeasible to calculate a reference total energy for limonin using CCSD(T) at an appropriately high basis set. Instead, we consider a reference total energy for limonin calculated at the MP2/cc-pCVQZ level of theory: approximately  $-1609.907 E_h$ . The atomization energy of limonin as per G4(MP2) was calculated as approximately  $11.183 E_h$ .

Per-iteration results for ALL-strategy adaptive calculations for limonin are shown in Figure 7, with a layout equivalent to those for heptane. We omit from consideration also the subgrid  $\Pi_{GCM}$ , since the results available up to practicality are not usefully comparable. We begin with approximations of the MP2/cc-pCVQZ total energy. In their final iterations, all such calculations, either approach ( $\Pi_{\text{non-ML}}$  and  $\Pi_{BS}$ ) or comfortably exceed ( $\Pi_{AbI}$  and  $\Pi$ ) chemical accuracy with respect to the reference result. Their respective error indicators are still sometimes prone to overestimation of the error, even late in the calculation. There does not seem to be either a clear speedup relative to the reference, or a clear winner amongst the subgrids trialled: the  $\Pi_{AbI}$  subgrid is the first to reach chemical accuracy, but the errors produced by truncations for  $\Pi$  seem to decay slightly more regularly.

For approximations of atomization energies, it is again interesting to see that while the error indicators are broadly reliable, they do underestimate the true error in the very final iterations. It is difficult to be conclusive here. As regards accuracy, neither calculation produces an agreement with the reference value better than  $0.01 E_h$ . Here, it may be the case that further iterations would produce a closer agreement, but again, since we would expect the truncations to approach the true  $V^{\text{atom}}$  — or, given the restricted grid, the approximation of  $V^{\text{atom}}$  for CCSD(T)/cc-pCV8Z — this would again depend on the accuracy of the G4(MP2) energy itself.

Although we are limited in our ability to draw conclusions by the quality of the available reference results, it does still seem clear that ML-SUPANOVA truncations are systematically improvable. It is important to note that the highly cyclic structure of the limonin system is not an issue. For atomization-energy calculations, there is still no clear performance benefit here compared to a simple application of G4(MP2). But the G4(MP2) cost is here significantly higher than that for heptane, and there must



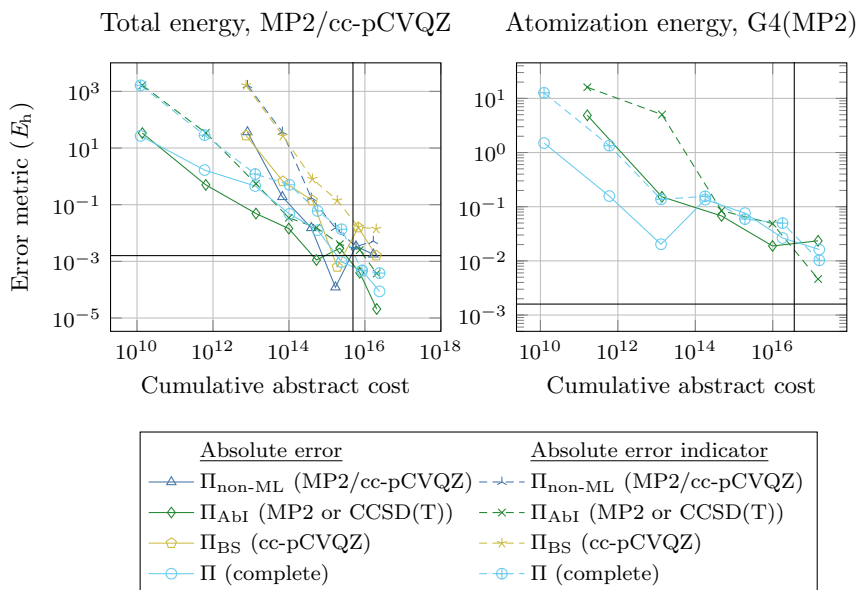


Figure 7: Per-iteration error metrics for adaptive truncations of (28) for limonin, as well as subgrid decompositions.

quickly come a point where even G4(MP2) will become prohibitively expensive. It is here that we still anticipate that the ML-SUPANOVA approach will truly come into its own.

## 5. Concluding Remarks

In this article, we investigated the working equations of single- and multilevel energy-based fragmentation methods. The technique of Möbius inversion, well-known to the cluster expansion methods community and easily extensible from work there, led to the unifying expansion form (23). As well as reproducing the energy equations of a number of existing multilevel fragmentation methods, this expansion provides for a further extension again, namely the efficient approximation of individual many-body terms in the broad style of a composite method, using a framework based on the combination technique.

The algorithm we have given here is clearly capable of refining ML-SUPANOVA truncations in a way that produces a systematic increase in accuracy. However, some questions do remain regarding their function as a mechanism for computational speedup. We can reasonably anticipate that application of ML-SUPANOVA is likely to pay dividends for much larger systems than those we have considered here. A further investigation of “true” adaptivity, with, e.g., the THRESHOLD strategy, would be appropriate. But the primary issue will remain one of validation, given the inherent unavailability of true reference solutions. Two complementary approaches suggest themselves. The first would be an effort to develop rigorous bounds on the decay properties of the MBE subproblem potentials  $\tilde{V}_{\mathbf{u}}$ , and then also on their more general multilevel equivalents. The second would be detailed practical comparison against experimental results. For this, it would be necessary to obtain from truncated ML-SUPANOVA expansions also point evaluations of the nuclear gradient, rather than just of total and/or atomization energies. This adjustment, very well-known in existing fragmentation methods, requires in our setup just a different choice of linear operator.

Were it not already clear, let us stress that the combinatorial tools we have applied here are textbook knowledge. Nevertheless, they allowed us to identify and avoid a subtle but important issue in the existing (ML-)BOSSANOVA framework. It seems likely that careful investigation of more recent developments in combinatorics, order theory, and the theory of graph convexities may lead to helpful further insights. It would be particularly interesting to explore whether and to what extent it might be possible to derive a general scheme for the construction of a interaction graph  $G$  that both captures the topological structure of a molecular system while also ensuring that  $\mathcal{M}_g[G]$  is truly a convex geometry, but this may prove a

difficult task.

As mentioned above, we remain particularly interested in the use of subproblem potentials employing quantum embedding, which are explicitly catered for by our formal setup. For example, if one were to build a set of potentials that applied either a WFT-in-WFT or WFT-in-DFT quantum embedding scheme, it would be natural to use the *ab initio* theory index  $m$  and basis set index  $p$  to specify the treatment of the embedding region. There are other possibilities; one might also consider a multilevel treatment of the environment region. Here, there will be practical issues, since the cost of evaluating each  $V_{(\mathbf{u},m,p)}$  will presumably scale superlinearly in the size of the full system.

## Acknowledgements

The authors gratefully acknowledge support from the Hausdorff Center for Mathematics (DFG project number 390685813) and the Collaborative Research Centre 1060 (DFG project number 211504053). The calculations reported were performed using the high-performance computing resources of Fraunhofer SCAI. We thank co-workers at the Institute for Numerical Simulation and at Fraunhofer SCAI, particularly Astrid Maaß, for many fruitful discussions. David Feller provided a helpful discussion regarding some of the basis sets used here.

## A. Proofs

In this section, we provide proofs for some of the results given in the text. With the exception of the first proof (for Prop. 2.1), these are drawn almost verbatim from [47], with some small adjustments for clarity and also consistency within this article. We use some basic terminology and concepts not explicitly introduced above; see, e.g., [30–32].

The first proof uses just the same standard results and idea of proof as on [35, p. 7481].

*Proof of Proposition 2.1.* If  $s \notin I$ , then immediately  $D_s^{(I)} = 0$ , so let  $s \in I$ . Define  $J := I \cup \{\hat{1}_J\}$ , such that  $\hat{1}_J >_J t$  for any  $t \in I$ . By Lemma 2.1,  $D_s^{(I)} = -\mu_J(s, \hat{1})$ . Note that the interval  $[s, \hat{1}]_J$  is itself a meet semilattice, so a lattice by [32, Prop. 3.3.1]. If  $\mu_J(s, \hat{1}) = -D_s^{(I)} \neq 0$ , then  $s$  must be a meet of coatoms of  $[s, \hat{1}]_J$  by [32, Cor. 3.9.5], and these coatoms are in  $A$  by construction.  $\square$

*Proof of Lemma 2.2.* Let  $P$ ,  $Q$ ,  $I'$ , and  $A'$  be as in the statement of the claim. Further, let  $I$  be some finite order ideal of  $P$  which is combination-consistent with  $I'$ . Since  $I$  is finite, it is generated by an antichain  $A \subseteq I$ ; we want to show that  $A = A'$ .

The proof is by contradiction. Suppose that  $A \neq A'$ . Then there exists some  $a \in A$  such that  $a \notin A'$ , or otherwise, some  $a' \in A'$  such that  $a' \notin A$ . Begin by supposing the former. Since  $a$  is maximal in  $I$ , we have  $D_a^{(I)} = \mu_P(a, a) = 1$  by definition. If  $a \notin Q$ , then by (17), also  $D_a^{(I)} = 0$ , a contradiction. So  $a \in Q$ . It must be that  $a \in I'$ , for otherwise,  $\hat{D}_a^{(I')} = 0$ , again a contradiction. Since  $a \notin A'$ , there exists some  $a^\dagger \in A'$  such that  $a^\dagger > a$ . Just as above,  $\hat{D}_{a^\dagger}^{(I')} = 1$ , but since  $a$  is maximal in  $I$ , we know that  $a^\dagger$  cannot also be an element of  $I$ , so also  $D_{a^\dagger}^{(I)} = 0$ , contradicting (17).

It must then be that there exists  $a' \in A'$  such that  $a' \notin A$ . Just as above,  $\hat{D}_{a'}^{(I')} = 1$ . If  $a' \notin I$ , then  $D_{a'}^{(I)} = 0$ , contradicting (17) since  $a' \in A' \subseteq Q$ . So  $a' \in I$ . Then there exists some  $a^\dagger \in A \subseteq P$  such that  $a^\dagger > a'$ , and  $D_{a^\dagger}^{(I)} = 1$ . Since this is nonzero, it must be that  $a^\dagger \in Q$ , by (17). But since  $\hat{D}_{a^\dagger}^{(I')}$  is nonzero only when  $a^\dagger \in I'$ , we have that  $a'$  is not maximal in  $I'$ , so  $a' \notin A'$ , a contradiction.  $\square$

In the following proof, we make use again of the same ideas as on [35, p. 7481], particularly for direction ( $\Leftarrow$ ). We correct here a minor omission in the proof given in [47], namely, a trivial case when  $s \notin I$ , and also correct a minor error related to the use of the Crosscut Theorem.

*Proof of Theorem 2.2.* ( $\Leftarrow$ ) Let  $P$  be a locally finite meet semilattice with a  $\hat{0}$ , and let  $Q \subseteq P$  be a meet subsemilattice of  $P$ . Let  $I'$  be an arbitrary finite order ideal of  $Q$  with generating antichain  $A'$ , and let  $I = \langle A' \rangle_P$  be the finite order ideal of  $P$  generated by  $A'$  in  $P$ .



Fix some  $s \in P$ . If  $s \notin I$ , then clearly  $D_s^{(I)} = 0$ , and if also  $s \in Q$ , similarly  $\hat{D}_s^{(I')} = 0$ . So assume  $s \in I$ . Define  $J := I \cup \{\hat{1}_J\}$  and  $J' := I' \cup \{\hat{1}_{J'}\}$ , such that  $\hat{1}_J >_J t$  for any  $t \in I$ , and  $\hat{1}_{J'} >_{J'} t'$  for any  $t' \in I'$ . If  $s$  is a meet of coatoms of  $J$ , then also  $s \in J'$ , since each such coatom is in  $A'$  by construction and  $Q$  is a meet subsemilattice of  $P$ . Further, noting that the coatoms of  $[s, \hat{1}_J]_J$  and  $[s, \hat{1}_{J'}]_{J'}$  must be the same, it follows that  $D_s^{(I)} = -\mu_J(s, \hat{1}_J) = -\mu_{J'}(s, \hat{1}_{J'}) = \hat{D}_s^{(I')}$  by Lemma 2.1 and the Crosscut Theorem [32, Cor. 3.9.4], using those sets of coatoms as the subset  $X$  in the statement of the latter result. If  $s$  is not a meet of coatoms of  $J$ , then  $D_s^{(I)} = 0$  by [32, Cor. 3.9.5]; if here also  $s \in Q$ , then similarly  $\hat{D}_s^{(I')} = 0$ , and we are done.

For  $(\Rightarrow)$ , the proof is by contradiction. Let  $P$  be a locally finite meet semilattice with a  $\hat{0}$ , and let  $Q$  be a combination-consistent subposet of  $P$ . Suppose that  $Q$  is, however, not a meet subsemilattice of  $P$ . Then there exist (at least) two distinct elements  $t, t' \in Q$  such that  $t \wedge_P t' \notin Q$ . Let  $I' = \langle \{t, t'\} \rangle_Q$  be the finite order ideal of  $Q$  generated by those two elements. Since  $Q$  is combination-consistent with  $P$ , there exists a finite order ideal  $I$  of  $P$  that is combination-consistent with  $I'$ . By Lemma 2.2, this  $I = \langle \{t, t'\} \rangle_P$ . Form as above  $J := I \cup \{\hat{1}_J\}$ . Since the only subset of  $\{t, t'\}$  whose meet is  $t \wedge_P t'$  is  $\{t, t'\}$  itself, it follows from the Crosscut Theorem [32, Cor. 3.9.4] that  $D_{t \wedge_P t'}^{(I)} = -\mu_J(t \wedge_P t', \hat{1}_J) = 1$ . But, since  $I$  is combination-consistent with  $I'$ , and since  $t \wedge_P t' \notin Q$ , also  $D_{t \wedge_P t'}^{(I)} = 0$ , a contradiction.  $\square$

## B. Abstract Cost Model

In this section, we outline the abstract cost model used in the calculations described above. The model is, in essence, just a collection of asymptotic expressions for the costs of individual algorithms, see again Section 2.1 and, e.g., [1–5], with the inclusion of some tunable constant factors. We simply state the model here, and refer the interested reader to [47] for more details as well as some additional supporting references.

The computational cost required to evaluate the total energy for a molecular system at some level of *ab initio* theory is ultimately a function of the specific molecular geometry as well as the selected basis set. The number and species of atoms determines the number of basis functions  $N_{\text{AO}}$ . All of the quantum-chemistry algorithms discussed here involve the calculation of formally  $\mathcal{O}(N_{\text{AO}}^4)$  two-electron integrals (ERIs) over atomic-orbital basis functions, but very many of these will usually be negligibly small [3, 49]. Since ERI negligibility is most commonly an expression of spatial separation between particles, this effect will be less pronounced for a more compact molecular system such as an individual fragment. In the following, we assume knowledge or at least an estimate of the number  $N_{\text{ERI}}$  of non-negligible ERIs. We also include a tuneable factor  $f_{\text{ERI}}$  that can be used to weight the cost of their calculation as  $f_{\text{ERI}}N_{\text{ERI}}$ .

We assume the use of a direct SCF algorithm [145] for Hartree-Fock calculations. The work at each iteration involves first calculating all distinct ERIs, and then solving the Roothaan-Hall equations. Assuming that some  $N_{\text{iter}}^{\text{HF}}$  iterations are required on average, then, the cost of a Hartree-Fock calculation is modelled as

$$\mathcal{C}^{\text{HF}}(\dots) = N_{\text{iter}}^{\text{HF}}(f_{\text{ERI}}N_{\text{ERI}} + N_{\text{AO}}^3),$$

cf. the cost model in [14]. We hide the various arguments and parameters of the cost function inside an ellipsis for notational simplicity.

It is common to see the cost of an MP2 calculation stated as  $\mathcal{O}(N_{\text{AO}}^5)$ , but this can be expressed a little more tightly. One necessarily begins with a Hartree-Fock calculation, which provides the first term in the following formula. The remaining calculation depends on the number of virtual orbitals  $N_{\text{virt}}$ , and also on the number of correlated orbitals  $N_{\text{corr}}$ , which is just the number of occupied orbitals  $N_{\text{occ}}$  for an all-electron calculation but somewhat lower for a calculation made under the frozen-core approximation. The actual evaluation of the MP2 contribution term requires only  $\mathcal{O}(N_{\text{corr}}^2 N_{\text{virt}}^2)$  operations, cf. [1], giving the second term below. However, these involve ERIs over molecular orbitals, which must be computed from the atomic-orbital ERIs. After these latter are explicitly calculated (third term), four tensor contractions (remaining terms) deliver the final tensor of molecular ERIs; for details on this transformation, see, e.g., [146]. Ordering these optimally under the assumption that  $N_{\text{corr}} < N_{\text{virt}}$ , then,

$$\begin{aligned} \mathcal{C}^{\text{MP2}}(\dots) = & \mathcal{C}^{\text{HF}}(\dots) + N_{\text{corr}}^2 N_{\text{virt}}^2 + f_{\text{ERI}} N_{\text{ERI}} \\ & + N_{\text{corr}} N_{\text{ERI}} + N_{\text{corr}}^2 N_{\text{AO}}^3 + N_{\text{corr}}^2 N_{\text{virt}} N_{\text{AO}} + N_{\text{corr}}^2 N_{\text{virt}}^2 N_{\text{AO}}. \end{aligned}$$

We model the costs of a general  $n$ th-order coupled cluster calculation more simply, as

$$\mathcal{C}^{\text{CC}(n)}(\dots) = f_{\text{ERI}} N_{\text{ERI}} + N_{\text{AO}}(N_{\text{corr}} + N_{\text{virt}})^4 + N_{\text{iter}}^{\text{CC}} N_{\text{corr}}^n N_{\text{virt}}^{n+2},$$

that is, as the cost of calculating non-negligible atomic ERIs (first term), transforming them into molecular ERIs (second term, modelled without as much detail as for MP2), and then iteratively solving the necessary amplitude equations, which is assumed to require  $N_{\text{iter}}^{\text{CC}}$  iterations on average; for asymptotic costs of coupled cluster calculations, see again, e.g., [8, 52, 53, 56]. This assumes that the cost difference of the frozen-core approximation can be captured by replacing  $N_{\text{occ}}$  by  $N_{\text{corr}}$ . Finally, by extension, the cost of evaluating such a calculation but also with a perturbative approximation for the effects of subsequent-order excitations is modelled as

$$\mathcal{C}^{\text{CC}(n)(n+1)}(\dots) = \mathcal{C}^{\text{CC}(n)}(\dots) + N_{\text{corr}}^{n+1} N_{\text{virt}}^{n+2}.$$

To obtain the abstract costs used in this article, we used the factor values  $N_{\text{iter}}^{\text{HF}} = 15$ ,  $N_{\text{iter}}^{\text{CC}} = 15$ , and  $f_{\text{ERI}} = 50$ . We used PySCF to derive values of  $N_{\text{AO}}$ ,  $N_{\text{corr}}$ , and  $N_{\text{virt}}$ , and estimated  $N_{\text{ERI}}$  as per the standard Cauchy-Schwarz bound [147], with the ERIs required to evaluate the bound calculated using LIBCINT [120]. ERIs were deemed negligible below a bound of  $10^{-12}$ , evaluated per basis shell, and counted up to permutational symmetry.

## C. Heuristic Fragmentation Algorithm

In this section, we describe informally the heuristic algorithm used to derive the fragmentations used in Section 4. Particularly similar heuristics are used in [16, 81]. As an initial candidate fragmentation, we take  $F = \{\{i\}\}_{i=1}^M$ , so assigning the index of each atom in the system to its own fragment. Fragments are then repeatedly coalesced in two separate phases.

The first phase collects any bonded hydrogen atoms with their “parents”, and removes the possibility that calculating any subproblem potential  $V_{F_u}$  would require severing a double- or higher-order bond. Iterating over pairs, two fragments  $F_i$  and  $F_j$  are combined if either a) a non-single covalent bond exists between an atom in each fragment, or b) a hydrogen atom in one fragment is bonded to an atom in the other fragment. This is repeated until a steady state is obtained.

The second phase is intended to avoid the possibility whereby two covalent bonds leading to the same external atom must be severed in the calculation of some  $V_{F_u}$ , which would result in the introduction of two link atoms in very close proximity to one another. We iterate again over pairs of fragments, looking for two particular cases. In the first case, a single atom in one fragment  $F_i$  is bonded to two atoms in another fragment  $F_j$ , where the two fragments are adjacent in the quotient graph. In the second case, atoms in two fragments  $F_i$  and  $F_j$  are bonded, and either of those two atoms is also bonded to an atom in a third fragment  $F_k$  that is itself adjacent to both  $F_i$  and  $F_j$ . Again, the second phase terminates when a steady state is encountered.

Note that the output of this algorithm is dependent upon the ordering of the fragments, and on how fragment pairs are enumerated at each iteration of each phase.

## References

- [1] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola, New York: Dover Publications, Inc., 1996.
- [2] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, et al. “Computational quantum chemistry: A primer”. In: *Special Volume, Computational Chemistry*. Vol. 10. Handbook of Numerical Analysis. Amsterdam: Elsevier, 2003, pp. 3–270.
- [3] P. Echenique and J. L. Alonso. “A mathematical and computational review of Hartree-Fock SCF methods in quantum chemistry”. *Mol. Phys.* 105.23-24 (2007), pp. 3057–3098.
- [4] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. Chichester: John Wiley & Sons Ltd., 2000.

- [5] F. Jensen. *Introduction to Computational Chemistry*. 3rd ed. Chichester: John Wiley & Sons, Ltd., 2017.
- [6] P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld. “Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited”. *J. Chem. Theory Comput.* 15.3 (2019), pp. 1546–1559.
- [7] K. Raghavachari and A. Saha. “Accurate composite and fragment-based quantum chemical models for large molecules”. *Chem. Rev.* 115.12 (2015), pp. 5643–5677.
- [8] A. Karton. “Quantum mechanical thermochemical predictions 100 years after the Schrödinger equation”. In: ed. by D. A. Dixon. Vol. 18. *Annual Reports in Computational Chemistry*. Elsevier, 2022, pp. 123–166.
- [9] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko. “Fragmentation methods: A route to accurate calculations on large systems”. *Chem. Rev.* 112.1 (2012), pp. 632–672.
- [10] M. A. Collins and R. P. A. Bettens. “Energy-based molecular fragmentation methods”. *Chem. Rev.* 115.12 (2015), pp. 5607–5642.
- [11] J. M. Herbert. “Fantasy versus reality in fragment-based quantum chemistry”. *J. Chem. Phys.* 151.17 (2019), p. 170901.
- [12] E. Suárez, N. Díaz, and D. Suárez. “Thermochemical fragment energy method for biomolecules: Application to a collagen model peptide”. *J. Chem. Theory Comput.* 5.6 (2009), pp. 1667–1679.
- [13] R. M. Richard and J. M. Herbert. “A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory”. *J. Chem. Phys.* 137.6 (2012), p. 064113.
- [14] S. R. Chinnamsetty, M. Griebel, and J. Hamaekers. “An adaptive multiscale approach for electronic structure methods”. *Multiscale Model. Simul.* 16.2 (2018), pp. 752–776.
- [15] R. M. Richard and J. M. Herbert. “Many-body expansion with overlapping fragments: Analysis of two approaches”. *J. Chem. Theory Comput.* 9.3 (2013), pp. 1408–1416.
- [16] H.-A. Le, H.-J. Tan, J. F. Ouyang, and R. P. A. Bettens. “Combined fragmentation method: A simple method for fragmentation of large molecules”. *J. Chem. Theory Comput.* 8.2 (2012), pp. 469–478.
- [17] S. N. Weiss, L. Huang, and L. Massa. “A generalized higher order kernel energy approximation method”. *J. Comput. Chem.* 31.16 (2010), pp. 2889–2899.
- [18] F. Heber. “Ein systematischer, linear skalierender Fragmentansatz für das Elektronenstrukturproblem”. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, 2014.
- [19] M. Griebel, J. Hamaekers, and F. Heber. “A bond order dissection ANOVA approach for efficient electronic structure calculations”. In: *Extraction of Quantifiable Information from Complex Systems*. Ed. by S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, et al. Vol. 102. *Lecture Notes in Computational Science and Engineering*. Cham: Springer International Publishing, 2014, pp. 211–235.
- [20] T. C. Ricard and S. S. Iyengar. “Efficient and accurate approach to estimate hybrid functional and large basis-set contributions to condensed-phase systems and molecule-surface interactions”. *J. Chem. Theory Comput.* 16.8 (2020), pp. 4790–4812.
- [21] J. H. Zhang, T. C. Ricard, C. Haycraft, and S. S. Iyengar. “Weighted-graph-theoretic methods for many-body corrections within ONIOM: Smooth AIMD and the role of high-order many-body terms”. *J. Chem. Theory Comput.* 17.5 (2021), pp. 2672–2690.
- [22] P. Seeber, S. Seidenath, J. Steinmetzer, and S. Gräfe. “Growing spicy ONIOMs: Extending and generalizing concepts of ONIOM and many body expansions”. *WIREs Comput. Mol. Sci.* 13.3 (2023), e1644.
- [23] E. E. Dahlke and D. G. Truhlar. “Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller-Plesset perturbation theory energies for large water clusters”. *J. Chem. Theory Comput.* 3.4 (2007), pp. 1342–1348.

- [24] G. J. O. Beran. “Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields”. *J. Chem. Phys.* 130.16 (2009), p. 164115.
- [25] J. Řezáč and D. R. Salahub. “Multilevel fragment-based approach (MFBA): A novel hybrid computational method for the study of large molecules”. *J. Chem. Theory Comput.* 6.1 (2010), pp. 91–99.
- [26] D. M. Bates, J. R. Smith, T. Janowski, and G. S. Tschumper. “Development of a 3-body:many-body integrated fragmentation method for weakly bound clusters and application to water clusters  $(\text{H}_2\text{O})_{n=3-10,16,17}$ ”. *J. Chem. Phys.* 135.4 (2011), p. 044123.
- [27] J. Hellmers and C. König. “A unified and flexible formulation of molecular fragmentation schemes”. *J. Chem. Phys.* 155.16 (2021), p. 164105.
- [28] R. M. Richard, K. U. Lao, and J. M. Herbert. “Understanding the many-body expansion for large systems. I. Precision considerations”. *J. Chem. Phys.* 141.1 (2014), p. 014108.
- [29] C. König and O. Christiansen. “Linear-scaling generation of potential energy surfaces using a double incremental expansion”. *J. Chem. Phys.* 145.6 (2016), p. 064105.
- [30] G.-C. Rota. “On the foundations of combinatorial theory I. Theory of Möbius functions”. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* 2.4 (1964), pp. 340–368.
- [31] M. Aigner. *Combinatorial Theory*. Classics in Mathematics. Berlin, Heidelberg: Springer, 1997.
- [32] R. P. Stanley. *Enumerative Combinatorics*. 2nd ed. Vol. 1. New York: Cambridge University Press, 2012.
- [33] R. Drautz, M. Fähnle, and J. M. Sanchez. “General relations between many-body potentials and cluster expansions in multicomponent systems”. *J. Phys.: Condens. Matter* 16.23 (2004), pp. 3843–3852.
- [34] D. J. Klein. “Chemical graph-theoretic cluster expansions”. *Int. J. Quantum Chem.* 30.S20 (1986), pp. 153–171.
- [35] L. Lafuente and J. A. Cuesta. “Cluster density functional theory for lattice models based on the theory of Möbius functions”. *J. Phys. A: Math. Gen.* 38.34 (2005), pp. 7461–7482.
- [36] M. Griebel, M. Schneider, and C. Zenger. “A combination technique for the solution of sparse grid problems”. In: *Iterative Methods in Linear Algebra*. Ed. by P. de Groen and R. Beauwens. North Holland: IMACS, Elsevier, 1992, pp. 263–281.
- [37] H.-J. Bungartz and M. Griebel. “Sparse grids”. *Acta Numer.* 13 (2004), pp. 147–269.
- [38] T. Gerstner and M. Griebel. “Dimension-adaptive tensor-product quadrature”. *Computing* 71.1 (2003), pp. 65–87.
- [39] M. Hegland. “Adaptive sparse grids”. *ANZIAM J.* 44 (2003), pp. 335–353.
- [40] M. Hegland, J. Garcke, and V. Challis. “The combination technique and some generalisations”. *Linear Algebra Appl.* 420.2-3 (2007), pp. 249–275.
- [41] B. Harding. “Adaptive sparse grids and extrapolation techniques”. In: *Sparse Grids and Applications — Stuttgart 2014*. Ed. by J. Garcke and D. Pflüger. Cham: Springer International Publishing, 2016, pp. 79–102.
- [42] M. Y. L. Wong. “Theory of the sparse grid combination technique”. PhD thesis. Australian National University, 2016.
- [43] M. Griebel. “Sparse grids for higher dimensional problems”. In: *Foundations of Computational Mathematics, Santander 2005*. Ed. by L. Pardo, A. Pinkus, E. Suli, and M. J. Todd. Cambridge: Cambridge University Press, 2006, pp. 106–161.
- [44] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. “On decompositions of multivariate functions”. *Math. Comput.* 79.270 (2009), pp. 953–966.
- [45] H. Rabitz and Ö. F. Aliş. “General foundations of high-dimensional model representations”. *J. Math. Chem.* 25.2/3 (1999), pp. 197–233.

- [46] J. Nieminen. “The lattice of connected subgraphs of a connected graph”. *Comment. Math.* 21.1 (1980), pp. 187–193.
- [47] J. N. J. Barker. “Order-Theoretic Combination Techniques and the Electronic Schrödinger Equation”. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, 2024.
- [48] R. Schneider. “Analysis of the projected coupled cluster method in electronic structure calculation”. *Numer. Math.* 113.3 (2009), pp. 433–471.
- [49] P. M. W. Gill. “Molecular integrals over Gaussian basis functions”. In: *Advances in Quantum Chemistry*. Ed. by J. R. Sabin and M. C. Zerner. Vol. 25. New York: Academic Press, 1994, pp. 141–205.
- [50] J. G. Hill. “Gaussian basis sets for molecular applications”. *Int. J. Quantum Chem.* 113.1 (2013), pp. 21–34.
- [51] C. Møller and M. S. Plesset. “Note on an approximation treatment for many-electron systems”. *Phys. Rev.* 46.7 (1934), pp. 618–622.
- [52] R. J. Bartlett and M. Musiał. “Coupled-cluster theory in quantum chemistry”. *Rev. Mod. Phys.* 79.1 (2007), pp. 291–352.
- [53] M. Kállay and P. R. Surján. “Higher excitations in coupled-cluster theory”. *J. Chem. Phys.* 115.7 (2001), pp. 2945–2954.
- [54] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon. “A fifth-order perturbation comparison of electron correlation theories”. *Chem. Phys. Lett.* 157.6 (1989), pp. 479–483.
- [55] Y. J. Bomble, J. F. Stanton, M. Kállay, and J. Gauss. “Coupled-cluster methods including noniterative corrections for quadruple excitations”. *J. Chem. Phys.* 123.5 (2005), p. 054101.
- [56] M. Kállay and J. Gauss. “Approximate treatment of higher excitations in coupled-cluster theory”. *J. Chem. Phys.* 123.21 (2005), p. 214105.
- [57] M. Kállay and J. Gauss. “Approximate treatment of higher excitations in coupled-cluster theory. II. Extension to general single-determinant reference functions and improved approaches for the canonical Hartree-Fock case”. *J. Chem. Phys.* 129.14 (2008), p. 144101.
- [58] T. H. Dunning Jr. “Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen”. *J. Chem. Phys.* 90.2 (1989), pp. 1007–1023.
- [59] D. Feller, K. A. Peterson, and J. G. Hill. “On the effectiveness of CCSD(T) complete basis set extrapolations for atomization energies”. *J. Chem. Phys.* 135.4 (2011), p. 044102.
- [60] L. A. Curtiss, P. C. Redfern, and K. Raghavachari. “Gaussian-4 theory”. *J. Chem. Phys.* 126.8 (2007), p. 084108.
- [61] L. A. Curtiss, P. C. Redfern, and K. Raghavachari. “Gaussian-4 theory using reduced order perturbation theory”. *J. Chem. Phys.* 127.12 (2007), p. 124105.
- [62] N. J. DeYonker, B. R. Wilson, A. W. Pierpont, T. R. Cundari, et al. “Towards the intrinsic error of the correlation consistent composite approach (ccCA)”. *Mol. Phys.* 107.8-12 (2009), pp. 1107–1121.
- [63] A. Karton, E. Rabinovich, J. M. L. Martin, and B. Ruscic. “W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions”. *J. Chem. Phys.* 125.14 (2006), p. 144108.
- [64] Y. J. Bomble, J. Vázquez, M. Kállay, C. Michauk, et al. “High-accuracy extrapolated ab initio thermochemistry. II. Minor improvements to the protocol and a vital simplification”. *J. Chem. Phys.* 125.6 (2006), p. 064108.
- [65] M. J. Field, P. A. Bash, and M. Karplus. “A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations”. *J. Comput. Chem.* 11.6 (1990), pp. 700–733.
- [66] J. Liu and J. M. Herbert. “Pair-pair approximation to the generalized many-body expansion: An alternative to the four-body expansion for ab initio prediction of protein energetics via molecular fragmentation”. *J. Chem. Theory Comput.* 12.2 (2016), pp. 572–584.

- [67] K. U. Lao, K.-Y. Liu, R. M. Richard, and J. M. Herbert. “Understanding the many-body expansion for large systems. II. Accuracy considerations”. *J. Chem. Phys.* 144.16 (2016), p. 164105.
- [68] J. F. Ouyang, M. W. Cvitkovic, and R. P. A. Bettens. “Trouble with the many-body expansion”. *J. Chem. Theory Comput.* 10.9 (2014), pp. 3699–3707.
- [69] J. Kongsted and O. Christiansen. “Automatic generation of force fields and property surfaces for use in variational vibrational calculations of anharmonic vibrational energies and zero-point vibrational averaged properties”. *J. Chem. Phys.* 125.12 (2006), p. 124108.
- [70] E. A. Bender and J. R. Goldman. “On the applications of Möbius inversion in combinatorial analysis”. *Amer. Math. Monthly* 82.8 (1975), pp. 789–803.
- [71] J. Quaintance and H. W. Gould. *Combinatorial Identities for Stirling Numbers. The Unpublished Notes of H. W. Gould*. Singapore: World Scientific Publishing Co Pte. Ltd., 2015. 284 pp.
- [72] L. O. Jones, M. A. Mosquera, G. C. Schatz, and M. A. Ratner. “Embedding methods for quantum chemistry: Applications from materials to life sciences”. *J. Am. Chem. Soc.* 142.7 (2020), pp. 3281–3295.
- [73] S. P. Veccham, J. Lee, and M. Head-Gordon. “Making many-body interactions nearly pairwise additive: The polarized many-body expansion approach”. *J. Chem. Phys.* 151.19 (2019), p. 194101.
- [74] D. Schmitt-Monreal and C. R. Jacob. “Frozen-density embedding-based many-body expansions”. *Int. J. Quantum Chem.* 120.21 (2020), e26228.
- [75] M. A. Collins and V. A. Deev. “Accuracy and efficiency of electronic energies from systematic molecular fragmentation”. *J. Chem. Phys.* 125.10 (2006), p. 104104.
- [76] R. Diestel. *Graph Theory*. 5th ed. Berlin: Springer, 2017.
- [77] F. Harary. *Graph Theory*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1972.
- [78] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. 4th ed. Cambridge, Massachusetts: The MIT Press, 2022.
- [79] T. C. Ricard, C. Haycraft, and S. S. Iyengar. “Adaptive, geometric networks for efficient coarse-grained ab initio molecular dynamics with post-Hartree-Fock accuracy”. *J. Chem. Theory Comput.* 14.6 (2018), pp. 2852–2866.
- [80] E. Prodan and W. Kohn. “Nearsightedness of electronic matter”. *Proc. Natl. Acad. Sci.* 102.33 (2005), pp. 11635–11638.
- [81] M. A. Collins. “Systematic fragmentation of large molecules by annihilation”. *Phys. Chem. Chem. Phys.* 14.21 (2012), pp. 7744–7751.
- [82] P. H. Edelman and R. E. Jamison. “The theory of convex geometries”. *Geom. Dedicata* 19.3 (1985).
- [83] I. M. Pelayo. *Geodesic Convexity in Graphs*. New York: Springer, 2013.
- [84] W. Goddard and O. R. Oellermann. “Distance in graphs”. In: *Structural Analysis of Complex Networks*. Ed. by M. Dehmer. Boston: Birkhäuser, 2011, pp. 49–72.
- [85] M. Farber and R. E. Jamison. “Convexity in graphs and hypergraphs”. *SIAM J. Algebraic Discrete Methods* 7.3 (1986), pp. 433–444.
- [86] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, et al. “ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(*t*-Bu)<sub>3</sub>)<sub>2</sub> + H<sub>2</sub> oxidative addition”. *J. Phys. Chem.* 100.50 (1996), pp. 19357–19363.
- [87] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, et al. “The ONIOM method and its applications”. *Chem. Rev.* 115.12 (2015), pp. 5678–5796.
- [88] N. J. Mayhall and K. Raghavachari. “Molecules-in-molecules: An extrapolated fragment-based approach for accurate calculations on large molecules and materials”. *J. Chem. Theory Comput.* 7.5 (2011), pp. 1336–1343.

- [89] D. R. Broderick and J. M. Herbert. “Scalable generalized screening for high-order terms in the many-body expansion: Algorithm, open-source implementation, and demonstration”. *J. Chem. Phys.* 159.17 (2023).
- [90] B. W. Bader and T. G. Kolda. “Efficient MATLAB computations with sparse and factored tensors”. *SIAM J. Sci. Comput.* 30.1 (2008), pp. 205–231.
- [91] Sparse developers. *Sparse*. Revision 94d196c3. 2018. URL: <https://sparse.pydata.org/en/stable/>.
- [92] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, et al. “Array programming with NumPy”. *Nature* 585.7825 (2020), pp. 357–362.
- [93] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, et al. “SciPy 1.0: Fundamental algorithms for scientific computing in Python”. *Nat. Methods* 17 (2020), pp. 261–272.
- [94] A. A. Hagberg, D. A. Schult, and P. J. Swart. “Exploring network structure, dynamics, and function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.
- [95] F. Johansson. “Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic”. *IEEE Trans. Comput.* 66.8 (2017), pp. 1281–1292.
- [96] P. Tol. *Qualitative colour schemes*. <https://personal.sron.nl/~pault/>. 2021.
- [97] J. Hansen. *Four Different Methods for Making Cel-Shaders in Blender Eevee (2.8, 2.9+)*. 2020. URL: <https://medium.com/@josephclaytonhansen/four-different-methods-for-making-cel-shaders-in-blender-eevee-2-8-2-9-6d976ce2555d> (visited on 10/22/2022).
- [98] J. Čížek. “On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods”. *J. Chem. Phys.* 45.11 (1966), pp. 4256–4266.
- [99] G. D. Purvis III and R. J. Bartlett. “A full coupled-cluster singles and doubles model: The inclusion of disconnected triples”. *J. Chem. Phys.* 76.4 (1982), pp. 1910–1918.
- [100] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. “Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields”. *J. Phys. Chem.* 98.45 (1994), pp. 11623–11627.
- [101] W. J. Hehre, R. F. Stewart, and J. A. Pople. “Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals”. *J. Chem. Phys.* 51.6 (1969), pp. 2657–2664.
- [102] W. J. Hehre, R. Ditchfield, R. F. Stewart, and J. A. Pople. “Self-consistent molecular orbital methods. IV. Use of Gaussian expansions of Slater-type orbitals. Extension to second-row molecules”. *J. Chem. Phys.* 52.5 (1970), pp. 2769–2773.
- [103] D. E. Woon and T. H. Dunning Jr. “Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon”. *J. Chem. Phys.* 98.2 (1993), pp. 1358–1371.
- [104] D. E. Woon and T. H. Dunning Jr. “Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon”. *J. Chem. Phys.* 103.11 (1995), pp. 4572–4585.
- [105] A. K. Wilson, T. van Mourik, and T. H. Dunning Jr. “Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon”. *J. Mol. Struct. THEOCHEM* 388 (1996), pp. 339–349.
- [106] K. A. Peterson, A. K. Wilson, D. E. Woon, and T. H. Dunning Jr. “Benchmark calculations with correlated molecular wave functions XII. Core correlation effects on the homonuclear diatomic molecules  $b_2-f_2$ ”. *Theor. Chim. Acta* 97.1-4 (1997), pp. 251–259.
- [107] D. Feller and K. A. Peterson. “Re-examination of atomization energies for the Gaussian-2 set of molecules”. *J. Chem. Phys.* 110.17 (1999), pp. 8384–8396.
- [108] S. L. Mielke, B. C. Garrett, and K. A. Peterson. “The utility of many-body decompositions for the accurate basis set extrapolation of ab initio data”. *J. Chem. Phys.* 111.9 (1999), pp. 3806–3811.
- [109] D. Feller and J. A. Sordo. “Performance of CCSDT for diatomic dissociation energies”. *J. Chem. Phys.* 113.2 (2000), pp. 485–493.

- [110] D. Feller and K. A. Peterson. “Probing the limits of accuracy in electronic structure calculations: Is theory capable of results uniformly better than “chemical accuracy”?” *J. Chem. Phys.* 126.11 (2007), p. 114105.
- [111] D. Feller, K. A. Peterson, and D. A. Dixon. “A survey of factors contributing to accurate theoretical predictions of atomization energies and molecular structures”. *J. Chem. Phys.* 129.20 (2008), p. 204105.
- [112] D. Feller and K. A. Peterson. “High level coupled cluster determination of the structure, frequencies, and heat of formation of water”. *J. Chem. Phys.* 131.15 (2009), p. 154306.
- [113] D. Feller, K. A. Peterson, and J. G. Hill. “Calibration study of the CCSD(T)-F12a/b methods for C<sub>2</sub> and small hydrocarbons”. *J. Chem. Phys.* 133.18 (2010), p. 184102.
- [114] J. H. Thorpe, J. L. Kilburn, D. Feller, P. B. Changala, et al. “Elaborated thermochemical treatment of HF, CO, N<sub>2</sub>, and H<sub>2</sub>O: Insight into HEAT and its extensions”. *J. Chem. Phys.* 155.18 (2021), p. 184109.
- [115] R. Ditchfield, W. J. Hehre, and J. A. Pople. “Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules”. *J. Chem. Phys.* 54.2 (1971), pp. 724–728.
- [116] W. J. Hehre, R. Ditchfield, and J. A. Pople. “Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules”. *J. Chem. Phys.* 56.5 (1972), pp. 2257–2261.
- [117] P. C. Hariharan and J. A. Pople. “The influence of polarization functions on molecular orbital hydrogenation energies”. *Theor. Chim. Acta* 28.3 (1973), pp. 213–222.
- [118] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, et al. “Gaussian-3 (G3) theory for molecules containing first and second-row atoms”. *J. Chem. Phys.* 109.18 (1998), pp. 7764–7776.
- [119] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, et al. “New basis set exchange: An open, up-to-date resource for the molecular sciences community”. *J. Chem. Inf. Model.* 59.11 (2019), pp. 4814–4820.
- [120] Q. Sun. “Libcint: An efficient general integral library for Gaussian basis functions”. *J. Comput. Chem.* 36.22 (2015), pp. 1664–1671.
- [121] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, et al. “PySCF: the Python-based simulations of chemistry framework”. *WIREs Comput. Mol. Sci.* 8.1 (2018).
- [122] Q. Sun, X. Zhang, S. Banerjee, P. Bao, et al. “Recent developments in the PySCF program package”. *J. Chem. Phys.* 153.2 (2020), p. 024109.
- [123] E. Aprà, E. J. Bylaska, W. A. de Jong, N. Govind, et al. “NWChem: Past, present, and future”. *J. Chem. Phys.* 152.18 (2020), p. 184102.
- [124] M. Kállay, P. R. Nagy, D. Mester, Z. Rolik, et al. “The MRCC program system: Accurate quantum chemistry from water to proteins”. *J. Chem. Phys.* 152.7 (2020), p. 074107.
- [125] MRCC, a quantum chemical program suite written by M. Kállay, P. R. Nagy, D. Mester, L. Gyevi-Nagy, J. Csóka, P. B. Szabó, Z. Rolik, G. Samu, J. Csontos, B. Hégyel, Á. Ganyecz, I. Ladjánszki, L. Szegedy, B. Ladóczki, K. Petrov, M. Farkas, P. D. Mezei, and R. A. Horváth. See [www.mrcc.hu](http://www.mrcc.hu).
- [126] Z. Rolik, L. Szegedy, I. Ladjánszki, B. Ladóczki, et al. “An efficient linear-scaling CCSD(T) method based on local natural orbitals”. *J. Chem. Phys.* 139.9 (2013), p. 094105.
- [127] M. Kállay. “A systematic way for the cost reduction of density fitting methods”. *J. Chem. Phys.* 141.24 (2014), p. 244113.
- [128] L. Gyevi-Nagy, M. Kállay, and P. R. Nagy. “Integral-direct and parallel implementation of the CCSD(T) method: Algorithmic developments and large-scale applications”. *J. Chem. Theory Comput.* 16.1 (2020), pp. 366–384.
- [129] F. D. Sönnichsen, C. I. DeLuca, P. L. Davies, and B. D. Sykes. “Refined solution structure of type III antifreeze protein: hydrophobic groups may be involved in the energetics of the protein-ice interaction”. *Structure* 4.11 (1996), pp. 1325–1337.



- [130] F. D. Sönnichsen, C. I. DeLuca, P. L. Davies, and B. D. Sykes. *North-Atlantic Ocean Pout Antifreeze Protein Type III Isoform HPLC12 Mutant, NMR, Minimized Average Structure*. Version 1.3; first published Apr. 1997. 2021. URL: <https://www.rcsb.org/structure/1kdf>.
- [131] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, et al. “Open Babel: An open chemical toolbox”. *J. Cheminf.* 3.1 (2011), p. 33.
- [132] T. K. Harris and A. S. Mildvan. “High-precision measurement of hydrogen bond lengths in proteins by nuclear magnetic resonance methods”. *Proteins: Struct., Funct., Genet.* 35.3 (1999), pp. 275–282.
- [133] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, et al. “Covalent radii revisited”. *Dalton Trans.* (21 2008), pp. 2832–2838.
- [134] A. George and J. W. H. Liu. “Algorithms for matrix partitioning and the numerical solution of finite element systems”. *SIAM J. Numer. Anal.* 15.2 (1978), pp. 297–327.
- [135] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, et al. “Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein”. *Cell* 181.2 (2020), pp. 281–292.
- [136] A. C. Walls, Y. J. Park, M. A. Tortorici, A. Wall, et al. *Structure of the SARS-CoV-2 spike glycoprotein (closed state)*. Version 2.1; first published Mar. 2020. 2021. URL: <https://www.rcsb.org/structure/6VXX>.
- [137] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer, et al. “Consistent van der Waals radii for the whole main group”. *J. Phys. Chem. A* 113.19 (2009), pp. 5806–5812.
- [138] K. Akisawa, R. Hatada, K. Okuwaki, Y. Mochizuki, et al. “Interaction analyses of SARS-CoV-2 spike protein based on fragment molecular orbital calculations”. *RSC Adv.* 11.6 (2021), pp. 3272–3279.
- [139] ChemSpider. *Heptane (CSID: 8560)*. Data as of 11/2022. URL: <https://www.chemspider.com/Chemical-Structure.8560.html>.
- [140] ChemSpider. *Limonin (CSID: 156367)*. Data as of 9/2021. URL: <http://www.chemspider.com/Chemical-Structure.156367.html>.
- [141] B. P. Pritchard and E. Chow. “Horizontal vectorization of electron repulsion integrals”. *J. Comput. Chem.* 37.28 (2016), pp. 2537–2546.
- [142] M. B. Ernst. *composite-thermochemistry-nwchem*. GitHub. 2016. URL: <https://github.com/mattbernst/composite-thermochemistry-nwchem>.
- [143] A. Karton, N. Sylvetsky, and J. M. L. Martin. “W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods”. *J. Comp. Chem.* 38.24 (2017), pp. 2063–2075.
- [144] S. K. Das, S. Chakraborty, and R. Ramakrishnan. “Critical benchmarking of popular composite thermochemistry models and density functional approximations on a probabilistically pruned benchmark dataset of formation enthalpies”. *J. Chem. Phys* 154.4 (2021), p. 044113.
- [145] J. Almlöf, K. Faegri Jr., and K. Korsell. “Principles for a direct SCF approach to LCAO-MO ab-initio calculations”. *J. Comput. Chem.* 3.3 (1982), pp. 385–399.
- [146] A. T. Wong, R. J. Harrison, and A. P. Rendell. “Parallel direct four-index transformations”. *Theor. Chim. Acta* 93.6 (1996), pp. 317–331.
- [147] M. Häser and R. Ahlrichs. “Improvements on the direct SCF method”. *J. Comput. Chem.* 10.1 (1989), pp. 104–111.