Journal of Computational Chemistry

WILEY

Journal of COMPUTATIONAL CHEMISTRY

RESEARCH ARTICLE OPEN ACCESS

Improving the Efficiency of Electrostatic Embedding Using the Fast Multipole Method

Pauline Colinet | Frank Neese | Benjamin Helmich-Paris

Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Germany

Correspondence: Pauline Colinet (colinet@kofo.mpg.de)

Received: 12 July 2024 | **Revised:** 31 October 2024 | **Accepted:** 6 November 2024

Funding: This work was supported by Deutsche Forschungsgemeinschaft, 388390466.

Keywords: ECM methods | FMM | ORCA

ABSTRACT

This paper reports the improvement in the efficiency of embedded-cluster model (ECM) calculations in ORCA thanks to the implementation of the fast multipole method. Our implementation is based on state-of-the-art algorithms and revisits certain aspects, such as efficiently and accurately handling the extent of atomic orbital shell pairs. This enables us to decompose near-field and far-field terms in what we believe is a simple and effective manner. The main result of this work is an acceleration of the evaluation of electrostatic potential integrals by at least one order of magnitude, and up to two orders of magnitude, while maintaining excellent accuracy (always better than the chemical accuracy of 1 kcal/mol). Moreover, the implementation is versatile enough to be used with molecular systems through QM/MM approaches. The code has been fully parallelized and is available in ORCA 6.0.

1 | Introduction

During the last three decades, electrostatic embedded-cluster methods (ECM) have proven very efficient for studying defects or local electronic properties in semi-conductors and insulators [1–8]. The strategy is to put most of the computational effort on a well localized area of interest, the finite cluster, using accurate but demanding quantum mechanical (QM) methods to describe it. The applied level of theory can be density functional theory (DFT), Hartree–Fock (HF), or higher-accuracy post-HF as offered by the employed QM calculator. The effect of the environment is retrieved in an approximate fashion by electrostatically embedding the cluster into an array of point charges (PC), the molecular mechanics (MM) area. In between the QM and MM areas, a layer of effective core pseudo-potentials (ECP) is inserted, which prevents the overpolarization of the QM density by the positive PCs in the MM part also known as the "electron spill-out" problem [9–12]. The PCs reproduce the long-range electrostatic interactions present in an infinite crystal. This can be achieved in two ways: either through a periodic array of PCs [13] or a finite array of PCs [14]. The unpolarized electrostatic potential generated by the MM PCs is incorporated into the one-electron part of the QM Hamiltonian. In the context of a QM/MM additive scheme, this is commonly referred to as the coupling term between the QM and MM regions. To refine this model, one can account for the induction effect by using a polarizable embedding, where the MM and QM components are mutually and self-consistently polarized during the QM calculations [15, 16].

One advantage of such ECM models is that they do not suffer from spurious self-interactions between defects or adsorbates that happen in the alternative periodic boundary conditions (PBC) calculations often used to study semi-conductors and insulators. One can indeed increase the size of the supercell in the PBC model to prevent such artificial interactions, but this comes with an increase of the computational cost. In this respect, the ECM models optimize computational resources (time and memory) by prioritizing the description of the electronic structure of

https://doi.org/10.1002/jcc.27532

the QM part and enable the use of broad range of methods from non-periodic QM programs, methods often not yet available under PBC [17, 18]. For example, the domain-based local pair natural orbital coupled-cluster theory (DLPNO-CCSD(T)) [19] is currently a reference method, providing access to more precise electronic energies than DFT approaches, with a linear scaling [20]. Both methods (ECM + DLPNO-CCSD(T)) have recently been used together to properly evaluate the adsorption energy of CO_2 on the surface of rutile (TiO_2) [21].

Interestingly, these ECM approaches also provide access to properties typically considered obtainable only through periodic calculations, such as the band gap of semi-conductors [22]. In some cases, ECMs can even yield better results than PBC calculations by accurately modeling the antiferromagnetic properties of oxide materials [23]. However, it is important to recognize that ECMs are only complementary to PBCs. While ECMs cannot be applied to metals with a vanishing band gap and lack the ability to describe critical aspects such as system symmetry and long-range dispersion corrections—which can be crucial depending on the system studied—they still offer valuable insights. When used alongside PBC calculations, ECMs enhance the overall understanding of the material properties.

Another field where those approaches could prove particularly useful is heterogeneous catalysis, where the cluster approach enables to access crucial information about the reaction intermediates [24]. In order to go beyond and reach real-surface electro-catalysis with accurate atomistic structure description under *operando* conditions, it is necessary to increase the cluster size. Today, it appears that the minimum number of PCs to reach the convergence of the property of interest (i.e., the calculated property is not impacted anymore by an increase of the size of MM area), can go up to $N_{PC} = 10^4$ if not 10^6 for QM clusters made of 200–300 atoms. However, for $10^4 - 10^5$ PCs, the evaluation of the PC electrostatic potential becomes one of the most demanding steps of the whole single-point QM calculation. It requires summing the Coulomb interactions between all relevant charge distributions within the system. Specifically, this process scales quadratically with the number of basis functions (N_{BF}) used to describe the quantum cluster, since interactions must be evaluated between each pair of basis functions. Additionally, the scaling is linear with respect to the number of point charges in the environment, as each point charge interacts with every basis function. Consequently, as the size of the system grows, both in terms of basis functions and point charges, the computational cost increases significantly, becoming more demanding for larger systems, with an overall scaling of $O(N_{PC} \times N_{BF}^2)$. As a matter of illustration, it already takes ≈ 30% of one single point calculation for a cluster of 136 NaCl atoms surrounded by 7×10^4 PCs at the DFT level of theory (PBE/def2-TZVP [25, 26]).

It would be beneficial to reduce the computation of this PC electrostatic potential, which serves as an approximate correction to account for the environment, in order to preserve one of the primary objective of ECM approaches, that is reduce the computational time attributed to the MM area. However, the Coulomb potential, $V(r) = \frac{1}{r}$, converges very slowly with distance r, so that long-range Coulombic interactions are difficult to ignore by truncation. On the quantum side, certain techniques like prescreening negligible integrals (e.g., the Schwartz inequality) can significantly reduce computational effort. In this context, we leveraged the highly efficient prescreening tools already implemented in ORCA [27]. However, the challenge remains in efficiently handling the classical environment's impact on the overall calculation. Established solutions to cope with this upscale effect of adding numerous PCs are the Ewald summation method with its improved particle mesh version [28, 29] and the fast multipole method (FMM) [30]. Both methods share the philosophy of splitting the interactions between a short and a long range terms, fixed by a distance parameter. In the short range (SR) regime, direct pairwise interactions are calculated explicitly, while in the long-range, the interactions are approximated in order to save computational time. The approximation, again, depends on a summation parameter. The Ewald summation methods have been developed to deal with periodic systems. There, the long-range interactions are evaluated through Fourier transform in the reciprocal space. On the contrary, the FMM can be used with both periodic [31–33] as well as finite embedding. Regarding the latter case, it has actually already been used and proven extremely useful in the context of QM/MM calculations, especially when more expensive polarizable embedding schemes are employed [34–40].

In this work we show how to improve the efficiency of ECM and (additive) QM/MM calculations using ORCA [1], by accelerating the evaluation of the unpolarized electrostatic potential generated by the MM PCs. Though many quantum chemistry packages are equipped with FMM or have been interfaced with FMM libraries [34, 37, 41–43], we believe that adding it to ORCA would be a great benefit to the community enabling to couple efficient ECM calculations with ORCA's cutting-edge electronic-structure methods. Therefore, we have implemented the very FMM (VFMM) in ORCA using state of the art algorithms based on an octree hierarchy, avoiding complex arithmetic and handling extent of atomic orbital shell pairs (SP) in an efficient and accurate way [44–47], for the evaluation of the PC electrostatic potential.

In the Theoretical Background section (Section 2), we introduce the algorithm and its relevant parameters necessary for the reader to understand the discussion. Details about the specific implementation in ORCA [48] are provided in Section 3. In a fourth section, we discuss the results obtained with the implemented FMM for a solid state calculation (cf. Section 4.1), both regarding the accuracy and the efficiency of the algorithm, and we provide suggestions for the selection of crucial model parameters. Following that, in Section 4.2, based on the conclusions drawn in Section 4.1, we apply the methodology to a biological example, photosystem II in a lipidic membrane, to ensure the implementation is general enough. The next section (Section 4.3) focuses on one of the main parameter (the Tree Depth [TD]) of the algorithm, offering a direct method for determining its value without the need for iterative adjustments. Finally, the last Section 4.4 provides a discussion on the scaling of the implemented method.

2 | Theoretical Background

In this section, we only reiterate the important considerations in order to make this paper self-contained and introduce the notations for the specific case of QM/PC interactions through FMM. Indeed, the FMM algorithm has already been fully described in several references [30, 37, 49].

The idea of the algorithm is handle short- (SR) and long-range (LR) interactions differently. In the SR or near-field (NF) regime, the electrostatic interaction between a point charge A ($q_A, \mathbf{r}_A$) and a Gaussian distribution overlap $\rho(\mathbf{r}) = \mu(\mathbf{r})\nu(\mathbf{r})$ with center in $\mathbf{r}_B$ ($\mathbf{r}_A \neq \mathbf{r}_B$), $U^{\mathbf{NF}}_{A\rho}$, can be expressed as:

$$U^{\mathbf{NF}}_{A\rho} = \int \frac{q_A \mu(\mathbf{r})\nu(\mathbf{r})}{|\,\mathbf{r}_A - \mathbf{r}\,|} d\mathbf{r} \tag{1}$$

In the paper the Coulomb constant, $k = \frac{1}{4\pi\epsilon_0}$, is set to 1, according to the atomic unit convention.

Since ORCA 6.0, these specific NF one-electron integrals are evaluated through SHARK which drives most integral-related tasks in ORCA [27].

In the LR or far-field (FF) regime, this Coulomb type interaction is approximated by a multipole expansion. In the following, we will introduce the main equations derived for the evaluation of this FF interaction, $U^{\mathbf{FF}}$ (more details on the derivation can be found in Supporting Information).

In a space of origin O(0,0,0), let A ($\mathbf{r}_A$) and B ($\mathbf{r}_B$) be the centers of two charge distributions (cf. Figure 1). The Coulomb interaction between these two centers can be expressed as:

$$U_{AB} = f\left(\frac{1}{|\mathbf{r}_A - \mathbf{r}_B|}\right) \tag{2}$$

The Coulomb interaction kernel $\frac{1}{|\mathbf{r}_A - \mathbf{r}_B|}$ can also be formulated by the partial-wave expansion:

$$\frac{1}{|\,\mathbf{r}_A - \mathbf{r}_B\,|} = \sum_{l=0}^{\infty} P_l(\cos(\theta)) \frac{\Delta r^l_{AB}}{\mathbf{r}^{l+1}_{QP}} \tag{3}$$

with $\Delta r_{AB} = \mathbf{r}_{AP} - \mathbf{r}_{BQ}$ and θ the angle between the two vectors $\mathbf{r}_{QP}$ and Δr_{AB}.

$P_l(\cos(\theta))$ are Legendre polynomials, with $|\,P_l(\cos(\theta))\,| \leq 1$, so that the series converge if and only if $|\,\Delta r_{AB}\,| < |\,\mathbf{r}_{QP}\,|$. This is the reason why A and B have to be well separated, which defines the FF regime.

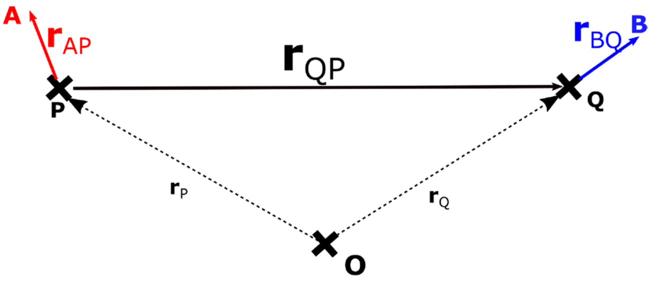


FIGURE 1 | Two particles A and B in the far field of each other, with their respective center of expansion P and Q.

Alternatively, following the work of Helgaker et al. the partial-wave expansion can be given in terms of regular ($R_{l,m}$) and irregular ($I_{l,m}$) solid scaled harmonics [46, 50]:

$$\frac{1}{|\,\mathbf{r}_A - \mathbf{r}_B\,|} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{j=0}^{\infty} \sum_{k=-j}^{j} (-1)^j R_{l,m}(\mathbf{r}_{AP}) I_{l+j,m+k}(\mathbf{r}_{QP}) R_{j,k}(\mathbf{r}_{BQ}) \tag{4}$$

with

$$R_{l,m}(\mathbf{r}) = \frac{1}{\sqrt{(l-m)!\,(l+m)!}} \mathbf{r}^l C_{l,m}(\mathbf{r}) \tag{5}$$

$$I_{l,m}(\mathbf{r}) = \sqrt{(l-m)!\,(l+m)!}\, \frac{C_{l,m}(\mathbf{r})}{\mathbf{r}^{l+1}} \tag{6}$$

and $C_{l,m}$, Racah normalized spherical harmonics.

We can now introduce the multipole moment $\mathbf{Q}(\mathbf{r}_C, \mathrm{P})$ expanded in P ($\mathbf{r}_P$) of a charge distribution centered in $\mathbf{r}_C$. It is a vector, with elements $Q_{l,m} = f\big(R_{l,m}(\mathbf{r} - \mathbf{r}_P)\big)$ defined by the pair of values (l, m).

In the case of a point charge A ($q_A, \mathbf{r}_A$) we have ($\mathbf{r} = \mathbf{r}_C$):

$$Q^A_{l,m}(\mathbf{r}_A, P) = q_A \times R_{l,m}(\mathbf{r}_A - \mathbf{r}_P) \tag{7}$$

For a Gaussian overlap distribution, $\mu\nu$ centered in $\mathbf{r}_B$ the multipole moment integrals are given by

$$Q^{\mu\nu}_{l,m}(\mathbf{r}_B, P) = \int \mu(\mathbf{r}')R_{l,m}(\mathbf{r}' - \mathbf{r}_P)\nu(\mathbf{r}')d\mathbf{r}' \tag{8}$$

Now the electrostatic interaction between $\mu(\mathbf{r})\nu(\mathbf{r})$ with center in $\mathbf{r}_B$ and a point charge A ($q_A, \mathbf{r}_A$) in its FF, can be expressed as:

$$\int \frac{q_A \mu(\mathbf{r})\nu(\mathbf{r})d\mathbf{r}}{|\,\mathbf{r}_A - \mathbf{r}\,|} = \mathbf{Q}^{\mu\nu}(\mathbf{r}_B, Q)\mathbf{I}(\mathbf{r}_{PQ})\mathbf{Q}^A(\mathbf{r}_A, P) \tag{9}$$

with Q and P being the respective centers of the multipole expansions and $\mathbf{I}$ being the interaction matrix which elements are defined as:

$$\mathbf{I}_{lm,jk}(\mathbf{r}_{PQ}) = (-1)^j I_{l+j,m+k}(\mathbf{r}_{PQ}) \tag{10}$$

In practice, the summation in Equation (4) is truncated to a maximum value of l, which is referred to as the maximum angular momentum ($\mathrm{L}_{\max}$) of the expansion. As stated in Ref. [46], one can actually estimate an upper boundary of the error, $\delta^{\mathrm{L}_{\max}}_{AB}$, due to the truncation of the expansion. Following Equation (3) and using summation formula on the hypergeometric series one gets:

$$\delta^{\mathrm{L}_{\max}}_{AB} = \sum_{l=0}^{\infty} \frac{\Delta^l r_{AB}}{r^{l+1}_{QP}} P_l(\cos\theta) - \sum_{l=0}^{\mathrm{L}_{\max}} \frac{\Delta^l r_{AB}}{r^{l+1}_{QP}} P_l(\cos\theta) \leq \sum_{l=0}^{\infty} \frac{\Delta^l r_{AB}}{r^{l+1}_{QP}} - \sum_{l=0}^{\mathrm{L}_{\max}} \frac{\Delta^l r_{AB}}{r^{l+1}_{QP}} \tag{11}$$

and an upper boundary for the error, $\delta^{\mathrm{L}_{\max},\max}_{AB}$, is then:

$$\delta_{AB}^{\mathrm{L_{max}},\max} = \sum_{l=0}^{\infty} \frac{\Delta^l r_{AB}}{r_{QP}^{l+1}} - \sum_{l=0}^{\mathrm{L_{max}}} \frac{\Delta^l r_{AB}}{r_{QP}^{l+1}} = \frac{1}{r_{QP} - \Delta r_{AB}} \left(\frac{\Delta r_{AB}}{r_{QP}} \right)^{\mathrm{L_{max}}+1} \tag{12}$$

From Equation (12), one can see that $\delta_{AB}^{\mathrm{L_{max}},\max}$ is a function of both the $\mathrm{L_{max}}$ and the r_{QP} distance between centers of expansion: the higher $\mathrm{L_{max}}$, the smaller the error; similarly, the bigger r_{QP}, the smaller the error.

The distinction between NF and FF regions arises from decomposing the space into discrete boxes. This process involves placing the entire system into a single box, which is then iteratively subdivided into eight smaller boxes, forming hierarchical levels. This hierarchical structure, known as an octree, is the natural outcome of domain decomposition in three dimensions (3D) (see Figure 2). The maximum level of subdivision, which indicates the depth of the tree, is referred to as the TD.

The NF is usually composed of one layer of 26 nearest neighbors boxes (in 3D) (dark purple boxes in Figure 2), while the FF contains ($8^{TD} - 27$) boxes. In our implementation we revisited the definition of the NF area (cf. Section 3), so that the number of NF boxes is always greater or equal than 27, and the QM distribution is well separated from the FF PCs.

For each box (B) at the deepest level (i.e., smallest boxes), containing elements of the QM subsystem, the electrostatic potential generated in B, $\mathbf{V}_{\mathrm{B}}$, due to PCs in the MM subsystem, can be divided between a close NF ($\mathbf{V}_{\mathrm{B}}^{\mathbf{NF}}$) and a distant FF potential ($\mathbf{V}_{\mathrm{B}}^{\mathbf{FF}}$):

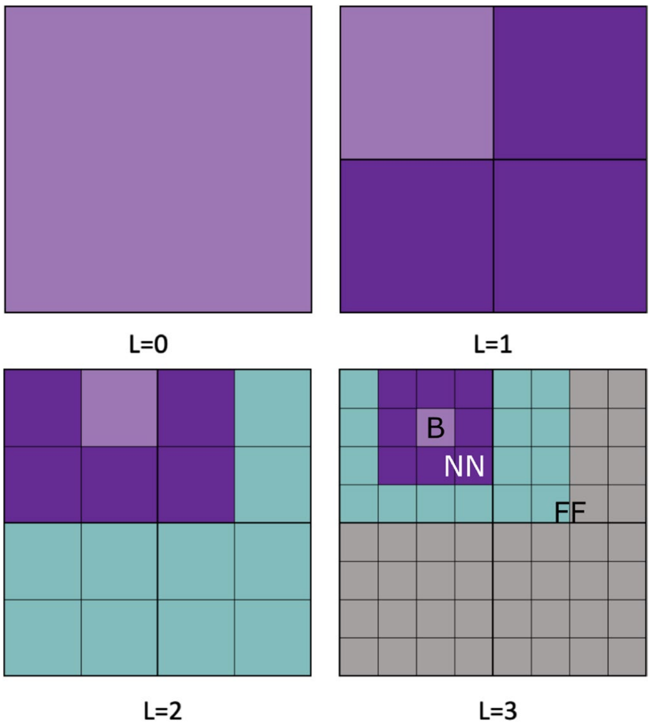


FIGURE 2 | Illustration of the octree box hierarchy in 2D. Up to three divisions of the initial box are shown, corresponding respectively to levels L = 1, L = 2, and L = 3. The nearest-neighbors boxes of a light purple box are represented with dark purple. The FF boxes ones are represented either with green when in the Local Far Field or with gray otherwise (Remote Far Field).

$$\forall \mathrm{B}, \quad \mathbf{V}_{\mathrm{B}} = \mathbf{V}_{\mathrm{B}}^{\mathbf{NF}} + \mathbf{V}_{\mathrm{B}}^{\mathbf{FF}} \tag{13}$$

with:

$$\mathbf{V}_{\mathrm{B}}^{\mathbf{NF}} = \sum_{A(\mathbf{r}_A) \in \mathrm{NF(B)}} \frac{q_A}{|\,\mathbf{r}_A - \mathbf{r}_B\,|} \tag{14}$$

Regarding the FF term, the center of the expansion (P) is arbitrary (cf. Equations (7) and (8)). Thus, for all PCs within a Box C centered in C, we can use C as center of expansion. By summing the resulting multipole expansions of each point charge within the box, we can form a single, unique multipole expansion:

$$\mathbf{Q}^{\mathrm{C}}(\mathbf{r}_C) = \sum_{A \in \mathrm{C}} \mathbf{Q}^A(\mathbf{r}_A, C) \tag{15}$$

and $\mathbf{V}_{B}^{\mathbf{FF}}$ becomes:

$$\mathbf{V}_{\mathrm{B}}^{\mathbf{FF}} = \sum_{\mathrm{C} \in \mathrm{FF(B)}} \mathbf{I}(\mathbf{r}_{CB}) \mathbf{Q}^{\mathrm{C}}(\mathbf{r}_C) \tag{16}$$

Besides, as visible in Figure 2, we can divide the FF area between a local FF (LFF) area and a remote FF area (RFF), so that a recursive scheme between levels appears:

$$\forall \mathrm{B}, \quad \mathbf{V}_{\mathrm{B}}^{\mathbf{FF}} = \mathbf{V}_{\mathrm{B}}^{\mathbf{LFF}} + \mathbf{V}_{\mathrm{B}}^{\mathbf{RFF}}, \tag{17a}$$

$$\mathbf{V}_{\mathrm{B}}^{\mathbf{FF}} = \mathbf{V}_{\mathrm{B}}^{\mathbf{LFF}} + \mathbf{W}^{T}_{Parent(_{\mathrm{B}}) \to _{\mathrm{B}}} \mathbf{V}^{\mathbf{FF}}_{Parent(_{\mathrm{B}})}, \tag{17b}$$

The LFF boxes are the boxes in the NF of the parent but not in the NF of the children, that represents a maximum of 189 boxes. The RFF potential is due to boxes which actually represent the FF area of the parent box of B. This means that one only needs to calculate the LFF term at every level, the rest of the potential will be inherited from the parent box. This is done by translating the center of the expansion of the potential from the center of the parent box to the center of Box B. The translation of these expansions is made possible by invoking again the addition theorem for solid harmonics, which enables to introduce the translation matrix $\mathbf{W}$ and its transposed form $\mathbf{W}^T$ [46]. The first one will be used to translate a multipole expansion ($\mathbf{Q}$) from one center to another (e.g., children to parent):

$$\mathbf{Q}(r_2) = \mathbf{W}(r_1 - r_2)\mathbf{Q}(r_1) \tag{18}$$

The second will be used to translate a potential ($\mathbf{V}_B = \mathbf{I}(\mathbf{r}_{PB})\mathbf{Q}_P$):

$$\mathbf{V}(r_2) = \mathbf{W}^T(r_1 - r_2)\mathbf{V}(r_1) \tag{19}$$

The elements of the translation matrix are defined as:

$$\mathbf{W}_{lm,jk}(\mathbf{r}_{PQ}) = R_{l-j,m-k}(-\mathbf{r}_{PQ}) \tag{20}$$

Finally, for every SP, the corresponding electrostatic perturbation to the one-electron Hamiltonian due to the PCs can be written as:

$$\mathbf{U}_{\mathrm{B}}^{\mu\nu} = \sum_{A(\mathbf{r}_A)\in \mathrm{NF(B)}} \int \frac{q_A \mu(\mathbf{r})\nu(\mathbf{r})d\mathbf{r}}{|\,\mathbf{r}_A - \mathbf{r}\,|} + \mathbf{Q}^{\mu\nu}\left(\mathbf{r}_{\mu\nu}, B\right)\mathbf{V}_{\mathrm{B}}^{\mathbf{FF}} \qquad (21)$$

3 | Implementation

The PC embedding FMM method was implemented in a development version of ORCA program package, and is now available within ORCA 6.0. With this version all calculations referred to in this paper were performed.

The algorithm can be summarized and divided in five principle steps:

1. Building the octree hierarchy, label QM boxes, and separate NF from FF boxes.
2. Evaluate the multipole moments at the deepest level for every box in the FF.
3. Translate the multipole moments from bottom to top of the tree (i.e., from children to parent).
4. Build $\mathbf{V}^{\mathbf{FF}}$ at the highest level and translate it to the bottom of the tree (i.e., from parent to children).
5. For every SP, build the corresponding multipole moment integrals (Equation 8) and evaluate the electrostatic potential integrals (Equation 21), which are added to the core Hamiltonian matrix (i.e., the one-electron part of the Fock matrix).

3.1 | Building the Octree Hierarchy

The total number of levels, which corresponds to the TD, is set by the user directly, or indirectly by setting the box dimension at the deepest level. In the latter case, the value of the box length at the deepest level, a, is doubled N times until the value obtained is greater or equal than the length of the initial box, a_0, surrounding the whole system: $2^N \times a \geq a_0$. The TD is then set accordingly to the value N.

An option has been implemented to refine the box dimension such that TD is kept unchanged but the dimension of the box is as small as possible (plus a small increment of 0.2 a.u. for safety). That is the box dimension is redefined according to:

$$a = \frac{a_0}{2^{\mathrm{TD}}} + 0.2 \qquad (22)$$

The impact of the optional refinement of a is discussed in 4.3. At every level, the boxes are indexed with three integers I, J, and K, defining their 3D coordinates (for instance the blue box in Figure 4 is defined by the indices (3,2,0)). These three integers are then zipped into a single integer, which was inspired from Gargantini's quadtree [45] and uniquely defines the box in the hierarchical grid. This unique identifier enables to efficiently access information such as the parent or children box index, and the box level. The zipping operation is done by combining the bits of I, J, and K integers in an interleaved manner bit by bit, using the AND, OR, and SHIFT operations. As a matter of illustration, the index of box ($I = 3$, $J = 2$, and $K = 0$) (at Level l = 2, i.e., third highest level) is obtained by combining the bits of 3 (**11**), 2 (**10**), and 0 (**00**) in the following way **011001** which is equal to $16 + 8 + 1 = \mathbf{25}$. Twenty five is the index of this specific box (3,2,0).

3.2 | Separation Between NF and FF (Well-Separateness Criterion)

Boxes containing SP centers are labeled as QM boxes, and all PCs within these boxes are, by definition, in the NF. The remaining boxes are then classified as either NF or FF.

For interacting discrete PCs, the difference between the distances of the PCs to their respective expansion centers (Δr_{AB}) must be smaller than the distance between their expansion centers ($\mathbf{r}_{QP}$) to guarantee convergence of the multipole expansion (cf. Equation (3)). Thus, it is enough to employ a single layer of NF boxes for an FMM implementation in the case of classical PC/PC interactions. However, for continuous charge distributions, which are not by definition confined to a single point in space, this is not true anymore. Hence, for every SP $\mu\nu$, we define a center of expansion, $P^{\mu\nu}$, and an extent parameter $\mathbf{r}_{ext}$, which delineate a sphere (centered in $P^{\mu\nu}$) of radius $\mathbf{r}_{ext}$ outside of which the overlap integrals and so the multipole integrals are negligible (i.e., $Q^{\mu\nu}\left(\mathbf{r}_{ext} - \mathbf{r}_P\right) < 10^{-5}$) (For more information see SI and Neese et al. [51]).

Then for any SP $\mu\nu$ in a Box B ($\mathbf{r}_B$) and centered in P ($\mathbf{r}_P$), the corresponding NF boxes are of two types: B's 26 nearest neighbors, and any Box C whose center is too close to the sphere of radius $\mathbf{r}_{ext}^{\mu\nu}$ centered in P. Mathematically, this condition is fulfilled whenever the distance $d = |\,\mathbf{r}_P - \mathbf{r}_C\,| - |\,\mathbf{r}_{ext}\,|$ is smaller than the diagonal of the box: $d < \sqrt{3} \times a$. If not, this ensures that no point in space is both within the SP sphere and Box C.

For efficiency reasons, we define a unique subset of NF boxes, containing all NF boxes of every SP of all QM boxes (cf. Figure 3). The PCs of this subset are then removed from the list of PCs and the electrostatic potential generated by them is calculated in a direct pairwise manner in ORCA's integral module SHARK. In that way we have just excluded all PCs that could be too close to the charge distribution for the multipole expansion to converge. All other PCs are expanded into multipole expansions and dealt with through the FMM routine.

3.3 | Evaluate the Multipole Moment of FF Boxes

The multipole moment generated by a point charge A (q_A, $\mathbf{r}$) is given by Equation (7), with $\mathbf{r}_P$ the center of the box. The regular solid scaled harmonics terms ($R_{l,m}$) are generated through a recursive scheme following Ref. [46]. The truncation parameter of the expansion ($\mathrm{L}_{\max}$) is set by the user. Boxes containing no particles are removed from the tree to gain efficiency. Complex numbers are avoided by handling real and imaginary components of R separately.

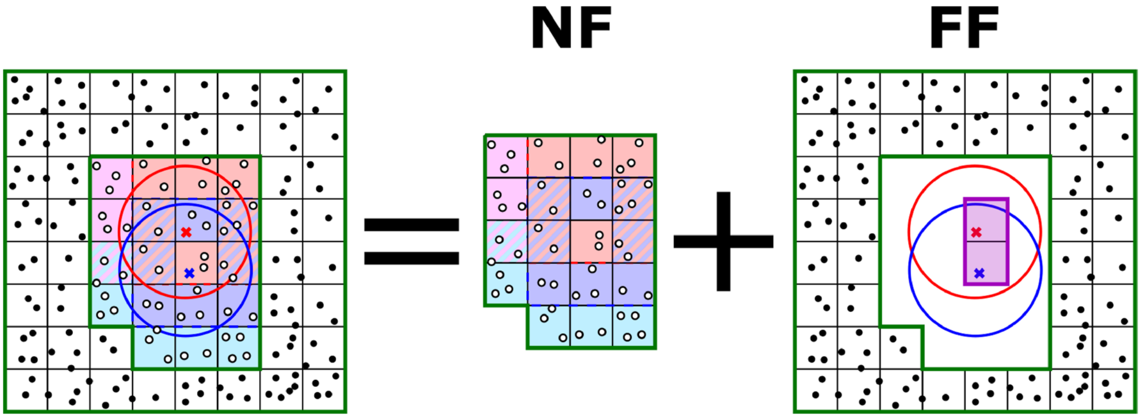


FIGURE 3 | Illustration of the separation between NF and FF. Two SP centers and their corresponding extent are shown respectively by red and blue crosses and red and blue circles of diameter r_{ext} (which can be different for every SP). The NF areas of the twos SPs are shown by light red and blue boxes, to which are added respectively the pink and cyan boxes after considering the extent of the SPs. The whole set of NF PC are then separated from the FF ones, the latter being dealt with through the FMM routine.

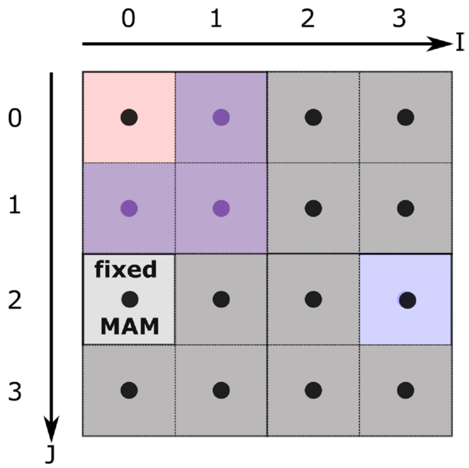


FIGURE 4 | Scheme of box indices in 2D to illustrate the VFMM strategy. The value K of the third dimension is 0 for every box shown here. Purple boxes indicate NF boxes of box ($I=0, J=0$, and $K=0$). Light gray box is the closest box used to define the reference maximum error to setup the VFMM.

3.4 | Evaluating Multipole Integrals

The multipole integrals over Gaussian SPs are currently evaluated in the Cartesian multipole basis through McMurchie Davidson scheme [52] and afterwards transformed into the spherical multipole basis using ORCA's integral module SHARK [28]. The multipole-moment integrals are always expanded in the box centers to avoid the costly origin translation of the integrals. The box in which the multipole moments integrals have their expansion center is assigned from the respective SP center (more details can be found in Ref. [51]).

3.5 | Translations

One can decompose the translation operations between two steps, the first one being the translation of the multipoles from the deepest to the highest level, the second one being the translation of the FF potential from the highest to the deepest level. Translation of the multipole moments from the children boxes to the parent boxes is done through the use of the **W** translation matrix. The translation of a potential from the top to the bottom of the tree is done using the transposed $\mathbf{W}^T$ matrix (cf. Equations (18) and (19)). The translation matrices are actually not build per se and stored. Instead, we calculate on the fly the non-zero elements of the matrix and contract them directly with either the multipole moments (Equation (18)) or with the FF potential vectors (Equation (18)). We use symmetry relationships between $R_{l,m}$ terms ($R_{l,-m} = (-1)^m R_{l,m}$) in order to optimize vectorization potential of the compiler.

As the value of L_{max} and, more importantly, the number of levels in the octree hierarchy increase, the cost of the translation operation increases as well. This is a subject that has been investigated in the past and for which solutions have been found, in particular the rotation around the z-axis which provides a scaling of the order $O(L^3_{max})$ instead of $O(L^4_{max})$ [53]. In the situation we discuss here, the evaluation of the electrostatic potential integrals through the FMM, the translation of the FF potential needs to be done only for boxes containing SP centers, that is QM boxes. By restricting this translation step to the QM boxes, the cost of the overall translation steps is dramatically decreased to the cost of the first translation step (i.e., translation of the multipole moments, cf. W_{MuMo} in Tables S2, S3, S4, and S5). As a result, translations are not the most limiting step here (cf. Supporting Information) and no more improvement of the efficiency of their evaluation has been considered for the moment.

3.6 | Building the FF Potential

Similar to the translation operations, the FF potential terms are computed on-the-fly to avoid storing large matrices [46]. This step involved the irregular solid scaled harmonics (Equation (10)), which are, as for the regular ones, evaluated through a recursive scheme. If the VFMM is turned on, this is where the truncation parameter is adapted according to the position of the box (cf. following paragraph).

3.7 | The “Very” Fast Multipole Method

Due to the recursive scheme between levels, the FF felt in the center of a box is built by only calculating interactions with the LFF boxes at every level (cf. Equations (17a) and (17b)).

While the evaluation of the FF potential, as previously mentioned, is not the most time consuming step, speeding up the LFF potential calculation at each step is still advantageous. Additionally, a straightforward method to achieve this has already been reported [44]. The key idea behind accelerating the LFF potential calculation is to use different $L_{\max}$ truncation parameters for the 189 boxes in the LFF area.

If we consider a Box B in (0,0,0), of dimension a, the closest boxes in the LFF are located at a distance $2a$. In order to make the discussion general for any level, we have divided the distance between the boxes by their dimension, a, such that in Figure 4, one of the closest LFF boxes is given by the position ($I=0$, $J=2$, and $K=0$). The error due to the truncation of the multipole expansion, $\delta^{\max}_{\text{closest box}}$, in this box can be evaluated using Equation (12), and considering the worst case scenario (wcs) for the position of PCs (definition of $\Delta r^{\text{wcs}}_{AB}$, which is a function of (I, J, and K), cf. Supporting Information). For any box in the LFF, we can similarly define $\Delta r^{\text{wcs}}_{AB}$, and r_{QP}, so that we can evaluate $\delta^{\max} < \delta^{\max}_{\text{closest box}}$. Then, the value of $L_{\max}$ in this box is decreased as long as the previous inequality is respected, with a minimum authorized value of $L_{\max}=2$.

All results shown and discussed in the following have been obtained using the VFMM option, unless otherwise stated.

4 | Results and Discussion

4.1 | Application to Solid State Calculations

In this section, we investigate the efficiency and accuracy of our newly implemented FMM PC embedding method for solid-state ECM calculations. The model considered has been used by one of us in the past to investigate the photochromic properties of sodalite mineral ($Na_8\left(AlSiO_4\right)_6Cl_2$) [4]. We have kept the same ECM protocol for sodalite using 93 atomic centers and 24 ECPs embedded in 45,883 PCs (Figure 5). The choice of the level of theory will obviously impact the overall calculation time. Here, we are running a single point DFT calculation with B3LYP hybrid-GGA [54–56] and double to triple-ζ split valence basis sets (see the reference paper [4] for more details). In this specific case then, the calculation of the one-electron part of the Hamiltonian takes 13% of the total time using analytic electrostatic potential (ESP) integrals for every environmental PC. With PBE [25] GGA functional, this goes up to 35% of the total time, using in both case the default convergence parameters for self-consistent field (SCF) calculations in ORCA 5.

4.1.1 | Accuracy

A faster algorithm is interesting if and only if it still provides good accuracy. That is why we first want to ensure that the approximation for the LR interactions, introduced by the FMM, does not spoil the overall accuracy. In that aim, we studied the impact of the different FMM parameters on the final single-point energy.

The reference energy E_0 is the one obtained without using the FMM for approximating the LR interactions. We note $\Delta E = E - E_0$, the error with respect to this reference (Figure 6).

The first parameter to set is the TD, that is, the number of times we divide the initial box. For the sodalite ECM calculation, the initial box has a dimension of 166.52 a.u. For TD = 2, the 64 boxes at the deepest level have a dimension of $a = 41.63$ a.u. (cf. Table 1).

The position and extent of some SPs is such that the well-separateness criterion is not met for all embedding boxes at this level. Hence, all PCs are considered to be in the NF of the QM boxes.

Thus the value of $L_{\max}$ will not change the final result, and the use of the FMM algorithm should not impact either the accuracy or the efficiency. However, in Figure S1, one can see that $\log_{10} | \Delta E |$ for TD = 2 is greater than the best numerical accuracy one can expect ($\Delta E = 10^{-16}$). This issue is related to the accumulation of numerical errors due to the finite precision of floating-point numbers. Consequently, the order of floating-point operations matters and can significantly impact the final result of a calculation: different sequences of operations can lead to different rounding errors. This phenomenon is known as numerical instability. For more detailed information on this topic, refer to the Standards for Floating-Point Arithmetic (IEEE 754) and related literature on numerical analysis [57].

As can be seen from Figure 6, the influence of the TD on the FMM error is negligible (not considering here the already discussed case of TD = 2). According to Equation (11), it is expected that the convergence of the energy as a function of $L_{\max}$ should be slower for smaller boxes (cf. SI “Discussion on the convergence speed”). This implies that the greater the distance between the QM boxes and their FF, the faster the convergence of the energy with increasing $L_{\max}$. However, in our implementation (cf. Section 3), the NF layer is defined to encompass the spheres of all SPs, so that the number of NF boxes is always greater than or equal to 27. Consequently, the distance between the QM boxes and the FF region remains similar regardless of TD value. Table 1, presents the total number of NF boxes, defining the NF at the deepest level for various TD values, along with the corresponding number of NF particles.

The second input parameter is the termination index for the multipole expansion, what we refer to as $L_{\max}$. As shown in Figure 6, the error decreases with the increase of the $L_{\max}$, in a non-linear way, reflecting indirectly the alternating signs in the partial-wave expansion values (cf. Equation (3)).

For an $L_{\max} \geq 15$, whatever the value of TD, one has reached an accuracy of 1 mHa (0.6 kcal/mol), which is lower than the “chemical accuracy” of 1 kcal/mol. If one wants to reach a μ Ha precision, an $L_{\max} \geq 23$ should be used.

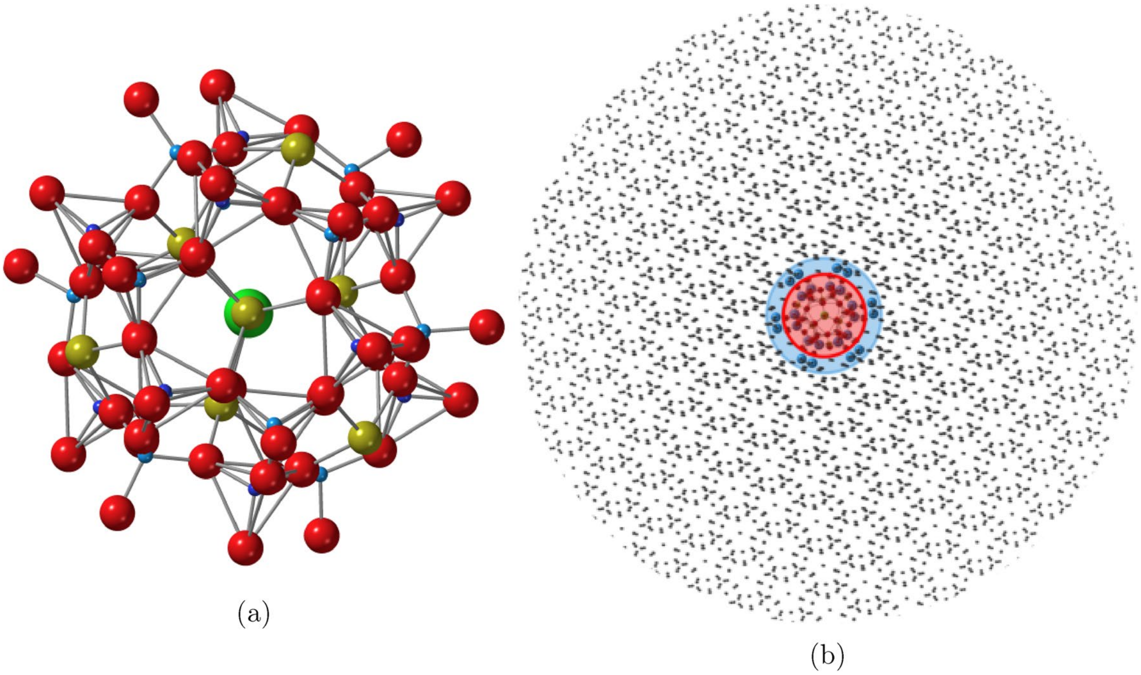


FIGURE 5 | (a) The QM cluster without ECPs—sodium, aluminum, silicon and oxygen atoms are respectively represented in yellow, cyan, blue, and red. The green sphere represents the F-center, responsible for the color of the material. (b) The QM cluster (red) surrounded by a layer of ECPs (light blue) and PCs (gray points).

TABLE 1 | Total number of boxes as well as box dimension and number of NF boxes at the deepest level for different TD input values, along with the corresponding number of NF particles at the deepest level.

TD	Number of boxes	Box dimension *a*	Number of NF boxes	Number of NF particles
2	64	41.63	64	45,883
3	512	20.92	136	12,113
4	4096	10.56	310	3397
5	32,768	5.38	1070	1569
6	262,144	2.79	3981	716

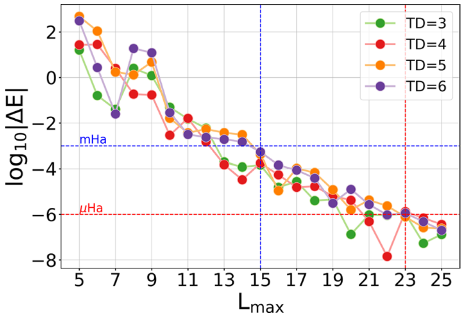


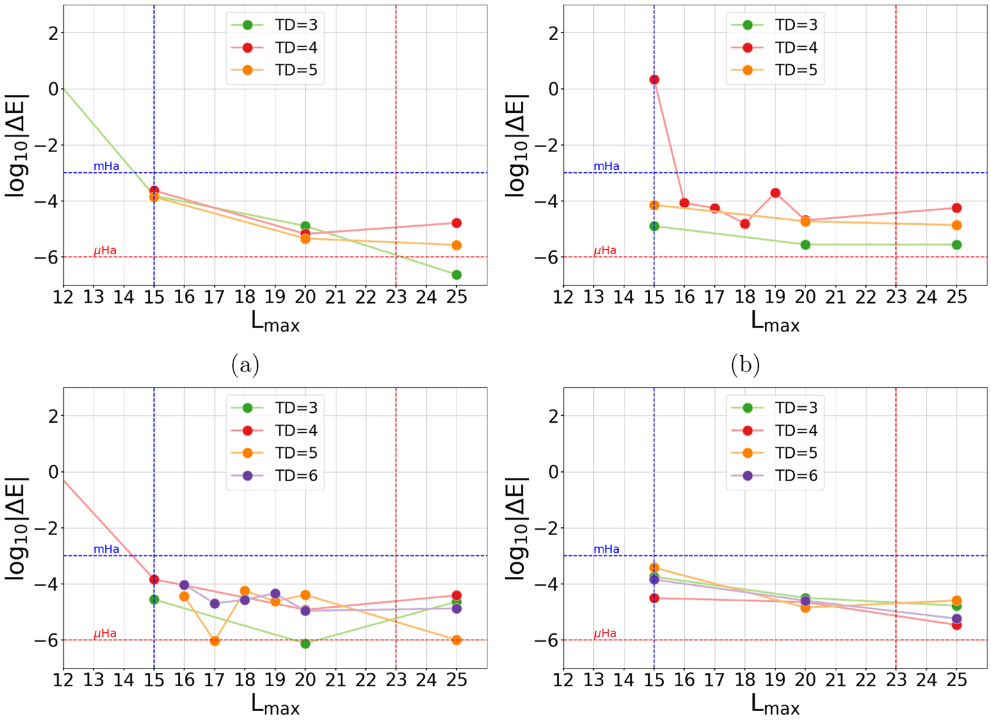
FIGURE 6 | Evolution of the error as a function of the maximum angular momentum (L_{max}) truncation parameter. Different TD has been used.

When the evaluation of the electrostatic embedding potential is actually far from the reference one (too big approximation due to too small L_{max}), this leads to a noticeable different starting point in the SCF (difference in the energy in the first step $>10^{-5}$ a.u.) which usually leads to an increased number of SCF iterations and large errors in the final SCF energy. These cases of deteriorated convergence only occurred for $L_{max} \leq 10$ (with or without VFMM). Thus, too small L_{max} should be avoided to prevent costly Fock matrix builds in the additional SCF cycles, overturning the overall efficiency of the implemented algorithm.

We tested the parameters on another (bigger) solid-state material example (NaCl), which was previously studied in the group, at PBE/def2-TZVP [25, 26] level of theory, to address the computation of solid-state chemical shifts in nuclear-magnetic resonance spectroscopy [5]. The clusters here are made of 136 QM atoms, 512 ECP, and PC embedding clusters of size 54,224, 73,440, 124,352, and 156,816. For all calculations we observe a good accuracy for $15 < L_{max} < 23$ (cf. Figure 7). An L_{max} of 20 would then be a safe choice without any further analysis of the system.

4.1.2 | Efficiency

Since the implementation leads to accurate results with an L_{max} high enough (≈ 20), we turn our attention now to the FMM efficiency. Again, the two parameters TD and L_{max} can impact the efficiency. In the following, the efficiency is assessed by looking



(a) (b) (c) (d)

FIGURE 7 | Evolution of the accuracy index parameter as a function of L_{max} and TD for NaCl system made of 136 QM atoms, 512 ECPs and respectively: (a) MM = 54,224 PCs, (b) MM = 73,440 PCs, (c) MM = 124,352, and (d) MM = 156,816 PCs.

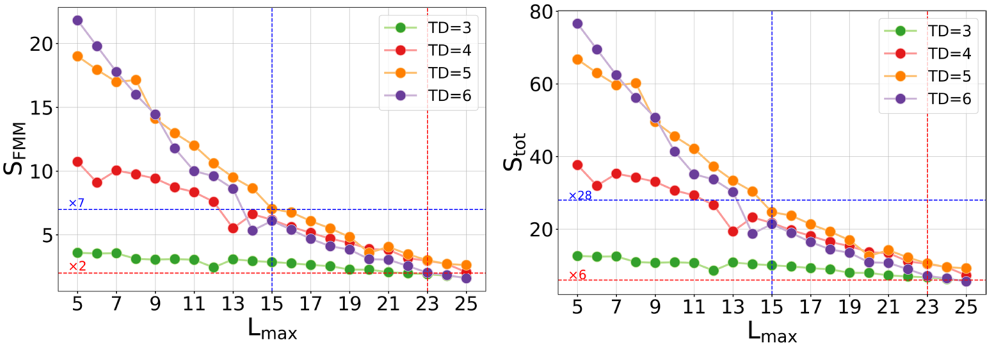
at the time required to evaluate the electrostatic potential integrals, using either analytic integrals for every PC (t_{ref}^{int}) or the implemented FMM version (t_{FMM}^{int}). The FMM speedup, S_{FMM}, is then defined as the ratio $S_{\text{FMM}} = t_{\text{ref}}^{int} / t_{\text{FMM}}^{int}$:

Additionally, we incorporated a newly optimized version of the electrostatic-potential (ESP) integrals code, built on the SHARK infrastructure in ORCA [27], which is now the default in ORCA 6. As a result, the evaluation of the NF one-electron integrals is also significantly accelerated compared to ORCA 5. The overall speedup, combining FMM with the faster NF one-electron integral evaluation, is referred to as S_{tot}. This leads to a dramatic improvement in performance, as shown in the case of sodalite for TD = 2, where no approximation in the FF is made: a 70% reduction in the time needed to evaluate the one-electron integrals is observed (192 s compared to 674 s, cf. SI Figure S4).

Finally, we examine the specific improvement introduced by the FMM (S_{FMM}) using the timing for TD = 2 as a reference. This reference represents the performance of ORCA 6 without the use of FMM (cf. Figure 8a). As anticipated, an increase in L_{max} corresponds to a decrease in efficiency. This is explained by the increasing cost as the truncation parameter of the multipolar expansion increases (cf. time for "Multipole integrals" in Figure 9). The evaluation of the multipole integrals, is obviously independent of TD.

Regarding the number of levels, we reach an optimum for TD = 5, with a systematic improvement from TD = 2 to TD = 5. This is due to the reduction of the number of PCs in the NF as we decrease the box dimension at the deepest level, illustrated by the reduction of the red boxes in Figure 9. At some point, the evaluation of the FF potential becomes the most expensive part of the calculation and the cost of adding more levels is visible. This is related to the computation and translation of the multipole expansions of the PCs in the FF and to a lesser extent to the translation of the FF potentials (cf. "Eval VFF" in Figure 9, and Tables S2, S3, S4, and S5). The cost becomes significant after six divisions of the initial box, making it less interesting to use TD = 6 than TD = 5. In the end, by using five levels, that is a box dimension of ≈ 5 a. u. here, and aiming at an accuracy of mHa (L_{max} = 15), one can speedup by a factor of eight the evaluation of the electrostatic potential. In total, the two new implementations (improved analytic integrals and FMM) lead in that case to a 28-times faster code (cf. Figure 8b). For a μ Ha accuracy, one can expect a 10-times faster calculation of the electrostatic potential from the embedding.

To illustrate the efficiency gained with the use of the VFMM option, we report in Figure 10 the best case with and without the use of VFMM, that is for TD = 5. In SI, one can find all other cases for different values of TD without using VFMM. Apart from values of $L_{\text{max}} \leq 10$, which have already been excluded from relevant choices, the use of VFMM as shown on Figure 10b provides similarly accurate results than the normal FMM algorithm even though the latter reaches the mHa accuracy for L_{max} = 14 instead of 15. Regarding the efficiency (cf. Figure 10a), the use of VFMM enables to reach slightly higher acceleration (S_{FMM}), from three to seven times faster ($15 \leq L_{\text{max}} \leq 23$) against two to six times faster without the use of the VFMM option.



(a)

(b)

FIGURE 8 | Speedup obtained for different values of TD, as a function of L_{max} (a) using ORCA 6 values as a reference (S_{FMM}) or (b) taking ORCA 5^1 values as a reference (S_{tot}).

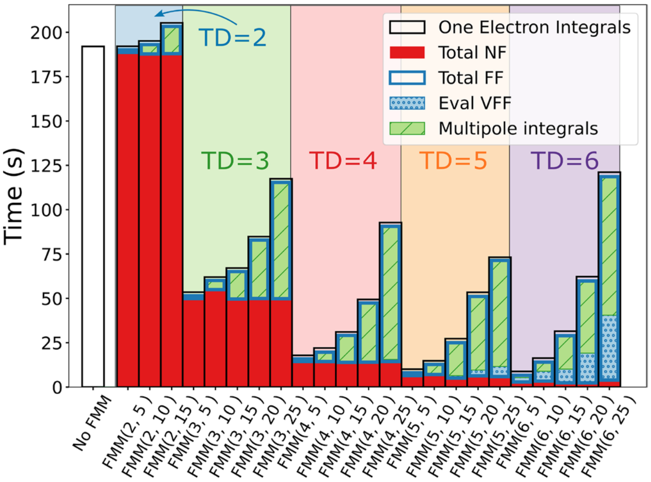


FIGURE 9 | Decomposition of the time spent in the calculation of the ESP integrals in ORCA for the sodalite test case system. The white bar (192 s) is obtained with ORCA 6 [1] without using the FMM. The same bar is then decomposed for different values of TD (blue = 2, green = 3, red = 4, yellow = 5, and purple = 6) and different values of L_{max} = 5, 10, 15, 20, or 25. The time is first decomposed into NF and FF contributions, respectively "Total NF" red filled box and "Total FF" blue stroke rectangle. The FF is further decomposed into the evaluation of the FF potential from the PCs (blue dots) and the evaluation of the multipole integrals from the SPs (green hatches). The remaining white space in the FMM bars is the time spent in the building of the octree.

In the end, if one wants to reach an accuracy of mHa, the use of TD = 5 and L_{max} = 15 with the VFMM will reduce to ≈ 0.5 % the time spent in the calculation of the electrostatic embedding potential along the whole SCF calculation (instead of 11 %). It would go up to ≈ 0.8 % for a more conservative L_{max} of 20. The impact of using the FMM is even more noticeable, for calculation with pure GGA for which no Hartree–Fock exchange is calculated. In the previously mentioned rock salt (NaCl) system, the time required for evaluating the electrostatic embedding potential with the PBE [25] functional accounts for 20.2% and 40.6% of the SCF calculation time for 54,224 and 156,816 PCs, respectively. In ORCA 6, without applying the FMM, this percentage decreases to 7% and 14%, respectively, due to an improved efficiency of the NF integral evaluation. When the VFMM is enabled with a TD of 5 and L_{max} set to 15, the evaluation time is further reduced to approximately 1% in both cases. However, if L_{max} is increased to 20, the time slightly increases to around 3%.

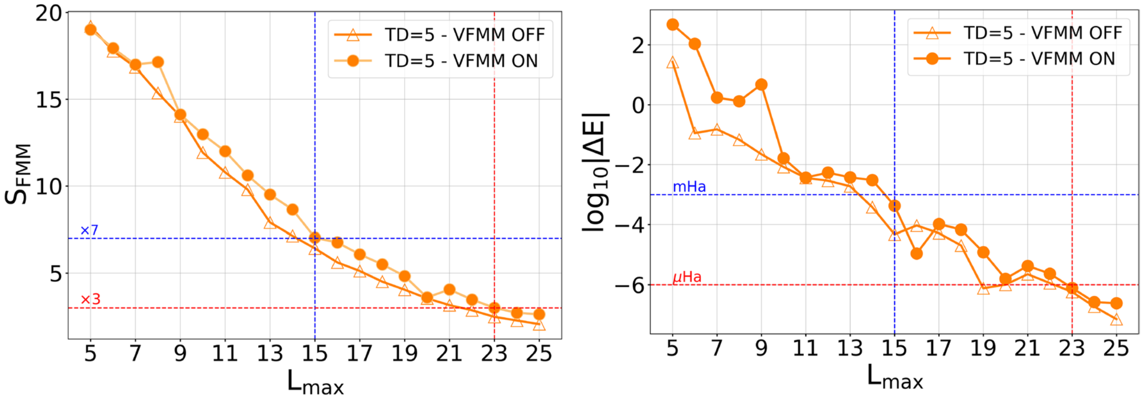
4.2 | Application to Biological Systems

In this section we investigate the performance of the implemented FMM for molecular QM/MM calculations with electrostatic embedding. From the perspective of the evaluation of the ESP integrals, such type of QM/MM calculations for molecules are identical to non-periodic ECM calculations for solids. The system taken as an example is the well-known photosystem II [58], with a setup used to investigate the nature and properties of triplet states within its reaction center [59, 60]. The QM part is made of two pigments of the system, with 153 atoms (cf. Figure 11b). The MM part is made of the rest of the protein system (75,894 atoms) and part of the lipid membrane (see for more details [60]), reaching a total of 503,195 atoms (cf. Figure 11a). The level of theory of the calculation is ωB97X-D/def2-TZVP [26, 61] with (D3BJ) dispersion correction [62].

For this system, the calculation of all one-electron integrals, including the evaluation of the electrostatic interaction from the MM embedding takes 28% of the single-point calculation in ORCA 5 (4 h and 28 min with one CPU core), 10% in ORCA 6.

We will use our new FMM implementation to reduce the calculation time. Based on the previous discussion on solid-state systems, we tested only a reduced number of values for L_{max}, all strictly greater than 10: 15, 20, and 25. If we look at the accuracy as a function of TD (cf. Figure 12a), as previously stated, it slightly decreases with addition of layers, but on the overall one can expect a μ Ha precision with L_{max} = 20 or 25. With L_{max} = 15 the accuracy is lower but still one to two orders of magnitude higher than the chemical accuracy. In terms of efficiency, the gain is obvious. For TD = 6 the evaluation of the electrostatic potential is accelerated 12, 20, and 31 times with the use of L_{max} = 25, 20, and 15, respectively (cf. Figure 12b). The total speedup with respect to ORCA 5 is respectively 46, 74, and 133 (cf. Supporting Information).

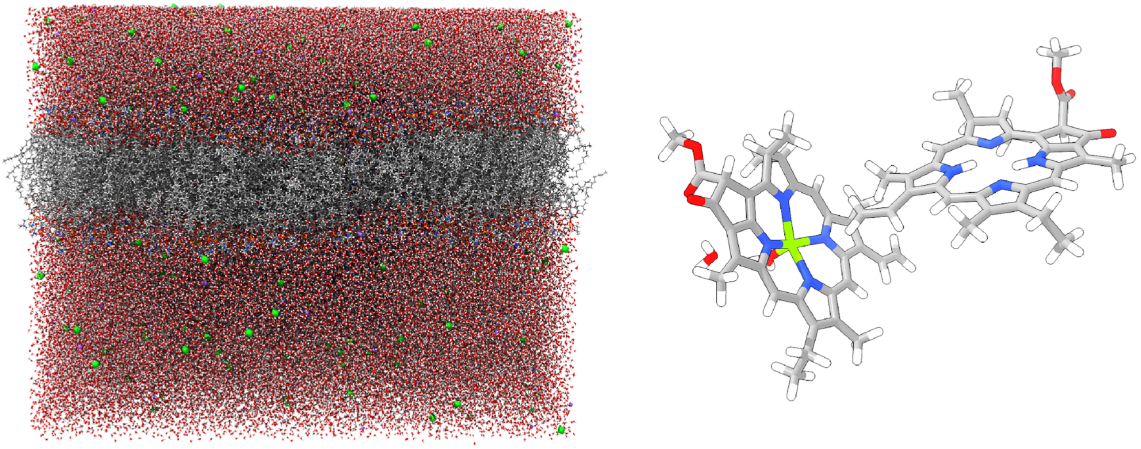
By adding another level using TD = 7, we start losing efficiency due to the cost of the additional translation of information in the tree. These extra costs are not balanced by the reduced time



(a)

(b)

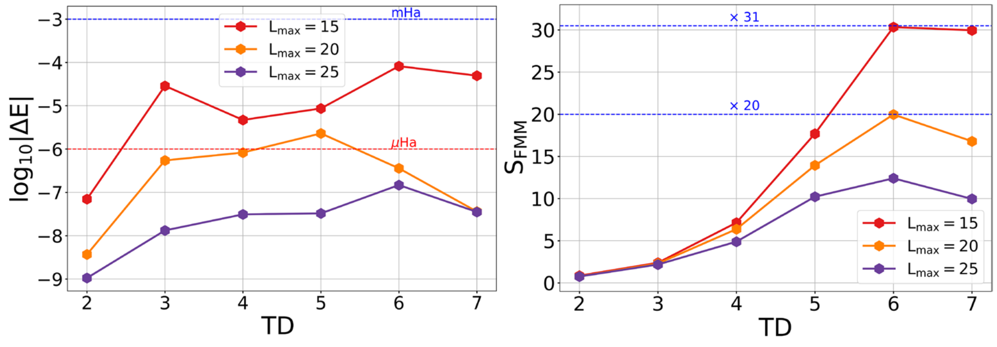
FIGURE 10 | Comparison between the FMM and VFMM algorithm in terms of (a) efficiency and (b) accuracy for the sodalite test case and TD = 5. The reference for the evaluation of both S_{FMM} and ΔE is the calculation made with ORCA 6 without using the FMM.



(a)

(b)

FIGURE 11 | (a) The whole QM/MM system including the membrane, water molecules, and the whole PSII. (b) QM atoms considered: C, N, O, H, and Mg represented respectively by gray, blue, red, white, and lime colors.



(a)

(b)

FIGURE 12 | (a) FMM accuracy for PSII test case as a function of the number of TD for different L_{max} values. (b) FMM speedup for PSII test case as a function of TD for different L_{max} values. $L_{max} = 15$ is in red, 20 in orange, and 25 in purple.

for the fewer NF interactions. Moreover, with TD = 7, the memory required for the calculation starts becoming a potential bottleneck. At Level 6, there are already $8^6 = 262{,}144$ boxes, at Level 7, we need 2,097,152. Storing the multipole expansion information for all boxes, with $L_{max} = 20$, requires large memory (48 GB for this system, cf. SI for more values).

4.3 | From TD to Box Dimension

Setting an appropriate TD parameter is less straightforward than setting the L_{max}. Increasing its value will not systematically lead to an increase in the accuracy. Setting TD to five works well in terms of accuracy and efficiency for the two test cases shown previously, sodalite and photosystem II, but for the latter TD = 6 gives a larger speedup and would be a better choice. If adding more levels when dealing with a bigger initial box is a priori the way to go, it is actually easier to think in terms of dimension of the box at the deepest level.

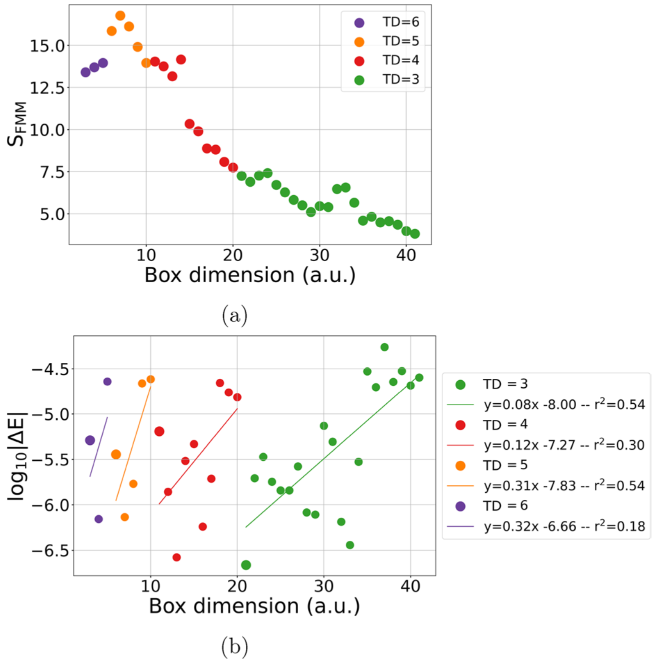
In the previous section, the dimension of the box at the deepest level a was set according to the number of levels (cf. Equation 22), with a_0 being the dimension of the initial cubic box surrounding the whole system, and 0.2 being a safety parameter to avoid particles being on the outside limit being neglected. The efficiency of our FMM implementation depends crucially on the box dimension a. Increasing a will reduce the time for evaluation of the FF potential $\mathbf{V}^{\mathbf{FF}}$, especially when this decreases the value of TD. On the contrary, increasing a increases the time for the evaluation of the NF potential $\mathbf{V}^{\mathbf{NF}}$. For $\mathbf{V}^{\mathbf{NF}}$, the time does not depend only on the number of layers but also and the extent of the SP overlap, which will impact the total number of NF boxes, Nb^{NF}. As a matter of illustration we report for different values of TD, the number Nb^{NF} when using either def2-SVP or def2-TZVP [26] basis set for the same PSII calculation. In the case of def2-TZVP, each atom type within the QM subsystem comes with more diffuse functions and thus SPs with greater extent, leading to an increased number of additional NF boxes (cf. Nb^{NF} values in Table 2). Interestingly, using $a = 6.464/\mathrm{TD} = 6$ leads to the best efficiency for both basis sets considered here (cf. Supporting Information).

Hence finding the optimal box dimension is not straightforward without testing several options. However, most systems, despite their extreme diversity, have a similar number of atoms per unit volume, suggesting that we can use the same division of space, that is, box dimensions, to obtain the optimal $Nb^{\mathrm{NF}}\,/\,Nb^{\mathrm{FF}}$ ratio. If we take a look at the two test cases again with parameters showing the best efficiency, the box dimension were respectively $a = 6.46$ Bohr for PSII with TD = 6, and $a = 5.38$ Bohr for sodalite with TD = 5, which are both close to 6 Bohr. Besides, if we imagine a system containing two times more particles than PSII and having similar density of particles, the volume of the initial box will be multiplied by two, but the dimension a_0 of the box will be only increased by 20% ($\times\, 2^{1/3}$), as will be a, without incrementing the number of levels.

Another important aspect is that for a given TD value, smaller boxes are better than larger boxes concerning both efficiency (cf. Figure 13a) and accuracy (cf. Figure 13b). Though this is more of a tendency (cf. r^2 values of the linear regression). Hence, the idea is to refine the box dimension to the smallest possible value while keeping the number of levels fixed (cf. Section 3). In that way, by starting from $a = 9$a.u. for both test cases, it will end up setting the box dimension to $a = 5.38$ and $a = 6.46$ respectively for sodalite and PSII. This value of $a = 9$a.u. should also cover similar systems with twice the number of particles since the value of a would then be $\approx 6.5 \times 2^{1/3} = 8.1$, without incrementing the number of levels.

TABLE 2 | Number of NF boxes Nb^{NF} and their dimension a at the deepest level for different TD and basis sets (def2-SVP or def2-TZVP [26]).

TD	a	def2-SVP	def2-TZVP
2	100.418	56	59
3	50.309	76	82
4	25.254	164	195
5	12.727	436	506
6	6.464	1357	1593
7	3.332	2999	3758



(a)

(b)

FIGURE 13 | Impact of the box dimension onto the accuracy and the efficiency, tested on the sodalite system. the corresponding values of TD are given thanks to a color code, that is respectively green, red, orange and purple for TD = 3, 4, 5, or 6. (a) Evolution of the efficiency ratio. (b) Evolution of the accuracy index with linear regression provided for every subset of values attributed to a specific value of TD.

In all tested cases (cf. SI), setting the initial box dimension to $a = 9$ Bohr will lead to the most or second most efficient parameter. The accuracy is mostly impacted by the L_{max} parameter. If one has to launch many calculations with the same setting, it is though recommended to launch a quick single point calculation with moderate value of L_{max} (e.g., 15) and different values of TD (e.g., 3, 4, 5, and 6) to determine the best settings. Still, the default parameters give excellent speedups, though improvements are possible when tuning them.

4.4 | Discussion on the Scaling of the Method

The evaluation of the electrostatic potential scales as $O(N_{PC}N_{BF}^2)$, with N_{PC} being the number of PCs in the environment and N_{BF} being the number of basis functions in the QM region. We believe that the number of basis functions is not the most relevant parameter to look at and we prefer to look instead at the number of remaining SPs after pre-screening [27]. The impact of the first term can be seen by increasing the number of PCs in the environment. This was investigated for NaCl rock salt system with 64 atoms in the QM region and different basis sets leading to different SP values: def2-SVP (SP = 63,160), def2-TZVP [26] (SP = 153,204), and def2-QZVP [63] (SP = 384,712). As expected, the time for the evaluation of the electrostatic integrals increases linearly with the number of PCs in the system without the use of the FMM (cf. Figure 14). Moreover, as the number of SPs in the system increases, the slope of this linear relationship steepens: the slope coefficients are respectively 0.0021, 0.0063, and 0.0230 for def2-SVP, def2-TZVP, and def2-QZVP (cf. Figure 14). However, when the FMM is employed, the slope decreases dramatically—by one to two orders of magnitude (e.g., 0.0001 against 0.0021 for def2-SVP, 0.0009 against 0.0230 for def2-QZVP). As a result, the impact of increasing the number of PCs becomes nearly negligible. The scaling of the method, which was previously linear with respect to the number of PCs, now depends mainly on the number of SPs, reducing it to a constant. For larger basis sets, it becomes advantageous to activate the FMM at a later stage compared to smaller sets, considering the number of PCs in the environment (cf. crossing of the red and pink curves on Figure 14).

In this initial study, Figure 14 shows that increasing the number of SPs impacts the timing, whether the FMM is used or not. This increase primarily affects the expansion of the multipole integrals, which is expected to scale linearly with the number of SPs and represents one of the most computationally expensive steps in the FMM process. It also influences the contraction with the FF potential during the construction of the Fock matrix. As a result, nearly linear scaling is anticipated when using the FMM.

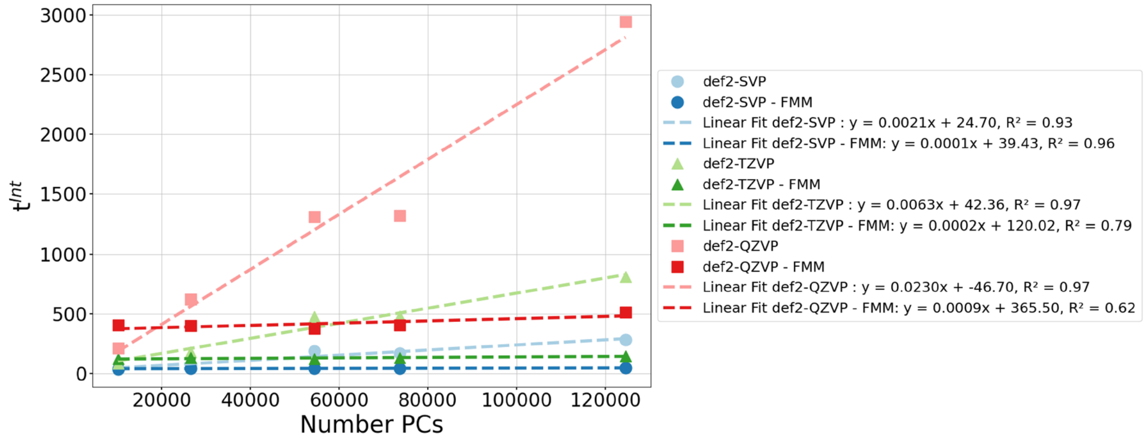


FIGURE 14 | Time required to evaluate the electrostatic-potential integrals as a function of the number of PCs in the environment, for NaCl system made of 64 QM atoms. Different basis sets have been considered: def2-SVP (blue curves), def2-TZVP (green curves) and def2-QZVP (red and pink curves). Darker colors stand for the use of the FMM, as also explicitly stated in the caption. Linear fits are reported with their equation and respective R^2 values in the caption.

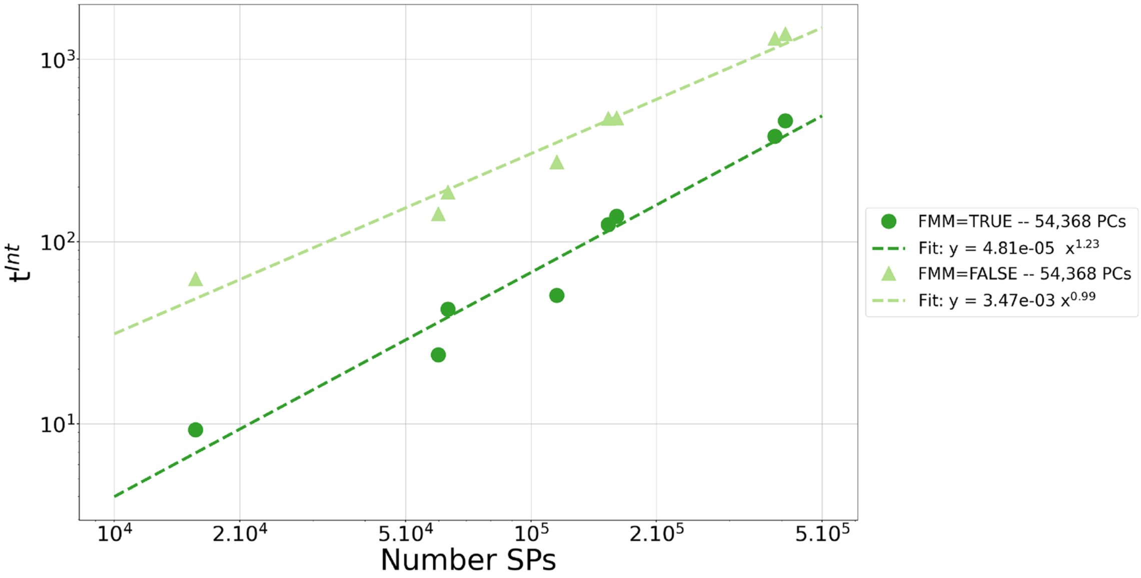


FIGURE 15 | Time required for the evaluation of the electrostatic-potential integrals as a function of the number of SPs in the QM part for NaCl system with 64 QM atoms and 54,368 PCs in the environment. Dark green data were obtained with the FMM (FMM = TRUE), and light green ones without (FMM = FALSE). Linear fits in the log scale are reported, with their respective equation in the caption.

To further investigate this, we fixed the number of PCs and varied the number of SPs by adjusting only the basis sets. We included def2-SVP, def2-TZVP, def2-QZVP, and added STO-3G (15,660 SPs), 3-21G (59,956 SPs), 6-311G (115,304 SPs), def2-TZVPP (160,380 SPs), and def2-QZVPP (407,536 SPs) in our study. Figure 15 reports results obtained with 54,368 PCs in the environment.

The data, presented on a log–log scale, were fitted linearly to extract the apparent scaling behavior of the method. Without the FMM (FMM = FALSE), the computational time scales nearly linearly with the number of point charges, with a scaling exponent of approximately 1 (×0.99). In contrast, when FMM is enabled (FMM = TRUE), the scaling exponent increases to just over 1.2, indicating a slightly super-linear scaling. This increase, compared to the ideal linear scaling, can be attributed to additional steps in the FMM algorithm, such as handling the expansion of quantum mechanical (QM) multipole integrals. However, the scaling remains far from quadratic, and more importantly, the computational cost when FMM = TRUE has a prefactor two orders of magnitude smaller than without FMM.

Thus, despite the slightly steeper scaling with the number of SPs when FMM = TRUE, the overall computational cost remains significantly lower compared to FMM = FALSE. Additionally, increasing the number of point charges (cf. Figure S10) has almost no impact on performance when FMM is enabled, while a consistent increase in runtime is observed when FMM = FALSE. This difference reflects the aforementioned linear dependence on the number of PCs in the non-FMM case, which disappears when the FMM is used.

As a matter of illustration, in the example shown on Figure 15, enabling FMM for systems with around 20,000 point charges consistently yields a performance benefit across the different basis sets tested.

5 | Conclusion

Based on previous developments, we provide an implementation of the FMM in ORCA for the evaluation of electrostatic potentials in the context of electrostatic embedding. We have shown that the implemented version of the FMM provides accurate results, easily above the chemical accuracy requested, while significantly speeding up the calculation. The tests indeed showed at least an order of magnitude increase in speed. The implementation is general and can be used in solid state as well as molecular applications. The PSII tests show that by using the FMM the size of the MM subsystem (with more than 500,000 atoms) is really not a limitation in terms of efficiency. Default parameters ($L_{max} = 20$, $a = 9.0$ a.u.) have been set to, first, ensure an accuracy in the total energy in the order of 0.1 mHa, and second, to provide faster calculations. Further investigations are currently conducted regarding the evaluation of the multipole integrals and will be reported in another article. We believe that our new implementation will be a major benefit, in particular when being applied in the field of heterogeneous catalysis. It offers an accurate description of the fundamental long-range Coulombic potential from the embedding at a limited computational cost. This efficiency allows for more precise study of the electronic structure of the quantum mechanical part using high-level methods for both ground-state and excited-state calculations.

Acknowledgments

The authors thank Dr. Anneke Dittmer for the rock salt structures, Dr. D. Pantazis, and S. Bhattacharjee for the PSII related inputs. The authors gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number 388390466–TRR 247, Project B8). Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. J. Vail, E. Emberly, T. Lu, M. Gu, and R. Pandey, “Simulation of Point Defects in High-Density Luminescent Crystals: Oxygen in Barium Fluoride,” *Physical Review B* 57 (1998): 764.

2. A. Kubas, M. Verkamp, J. Vura-Weis, F. Neese, and D. Maganas, “Restricted Open-Shell Configuration Interaction Singles Study on M- and L-Edge X-Ray Absorption Spectroscopy of Solid Chemical Systems,” *Journal of Chemical Theory and Computation* 14 (2018): 4320–4334.

3. T. Biswas and M. Jain, “Electronic Structure and Optical Properties of F Centers in α-Alumina,” *Physical Review B* 99 (2019): 144102.

4. P. Colinet, A. Gheeraert, A. Curutchet, and T. Le Bahers, “On the Spectroscopic Modeling of Localized Defects in Sodalites by TD-DFT,” *Journal of Physical Chemistry C* 124 (2020): 8949–8957.

5. A. Dittmer, G. L. Stoychev, D. Maganas, A. A. Auer, and F. Neese, “Computation of NMR Shielding Constants for Solids Using an Embedded Cluster Approach With DFT, Double-Hybrid DFT, and MP2,” *Journal of Chemical Theory and Computation* 16 (2020): 6950–6967.

6. R. Shafei, D. Maganas, P. J. Strobel, P. J. Schmidt, W. Schnick, and F. Neese, “Electronic and Optical Properties of Eu^{2+}-Activated Narrow-Band Phosphors for Phosphor-Converted Light-Emitting Diode Applications: Insights From a Theoretical Spectroscopy Perspective,” *Journal of the American Chemical Society* 144 (2022): 8038–8053.

7. R. Shafei, P. J. Strobel, P. J. Schmidt, D. Maganas, W. Schnick, and F. Neese, “A Theoretical Spectroscopy Study of the Photoluminescence Properties of Narrow Band Eu^{2+}-doped Phosphors Containing Multiple Candidate Doping Centers. Prediction of an Unprecedented Narrow Band Red Phosphor,” *Physical Chemistry Chemical Physics* 26 (2024): 6277–6291.

8. D. Berger, A. J. Logsdail, H. Oberhofer, et al., “Embedded-Cluster Calculations in a Numeric Atomic Orbital Density-Functional Theory Framework,” *Journal of Chemical Physics* 141 (2014): 024105.

9. P. J. Hay and W. R. Wadt, “Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for K to Au Including the Outermost Core Orbitals,” *Journal of Chemical Physics* 82 (1985): 299–310.

10. L. Kantorovich, “An Embedded-Molecular-Cluster Method for Calculating the Electronic Structure of Point Defects in Non-Metallic Crystals. I. General Theory,” *Journal of Physics C: Solid State Physics* 21 (1988): 5041.

11. P. Slavíček and T. J. Martínez, “Multicentered Valence Electron Effective Potentials: A Solution to the Link Atom Problem for Ground and Excited Electronic States,” *Journal of Chemical Physics* 124 (2006): 084107.

12. G. A. DiLabio, M. M. Hurley, and P. A. Christiansen, “Simple One-Electron Quantum Capping Potentials for Use in Hybrid QM/MM

Studies of Biological Molecules," *Journal of Chemical Physics* 116 (2002): 9578–9584.

13. A. M. Burow, M. Sierka, J. Döbler, and J. Sauer, "Point Defects in CaF_2 and CeO_2 Investigated by the Periodic Electrostatic Embedded Cluster Method," *Journal of Chemical Physics* 130 (2009): 174710.

14. M. Klintenberg, S. Derenzo, and M. Weber, "Accurate Crystal Fields for Embedded Cluster Calculations," *Computer Physics Communications* 131 (2000): 120–128.

15. J. M. Olsen, K. Aidas, and J. Kongsted, "Excited States in Solution Through Polarizable Embedding," *Journal of Chemical Theory and Computation* 6 (2010): 3721–3734.

16. J. M. H. Olsen and J. Kongsted, "Molecular Properties Through Polarizable Embedding," in *Advances in Quantum Chemistry*, vol. 61, eds. J. R. Sabin and E. J. Brändas (Amsterdam, Netherlands: Elsevier, 2011).

17. I. Y. Zhang and A. Grüneis, "Coupled Cluster Theory in Materials Science," *Frontiers in Materials* 6 (2019): 123.

18. B. X. Shi, A. Zen, V. Kapil, P. R. Nagy, A. Grüneis, and A. Michaelides, "Many-Body Methods for Surface Chemistry Come of Age: Achieving Consensus With Experiments," *Journal of the American Chemical Society* 145 (2023): 25372–25381.

19. C. Riplinger and F. Neese, "An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method," *Journal of Chemical Physics* 138 (2013): 034106.

20. Y. Guo, C. Riplinger, U. Becker, et al., "An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local Pair-Natural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD (T)]," *Journal of Chemical Physics* 148 (2018): 011101.

21. A. Kubas, D. Berger, H. Oberhofer, D. Maganas, K. Reuter, and F. Neese, "Surface Adsorption Energetics Studied With "Gold Standard" Wave-Function-Based Ab Initio Methods: Small-Molecule Binding to TiO_2(110)," *Journal of Physical Chemistry Letters* 7 (2016): 4207–4212.

22. A. Dittmer, R. Izsak, F. Neese, and D. Maganas, "Accurate Band Gap Predictions of Semiconductors in the Framework of the Similarity Transformed Equation of Motion Coupled Cluster Theory," *Inorganic Chemistry* 58 (2019): 9303–9315.

23. A. Dittmer, "Exploring Problems in Inorganic Solid-State Systems With Wavefunction-Based Molecular Spectroscopy Methods" (thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2024).

24. G. A. Bramley, O. T. Beynon, P. V. Stishenko, and A. J. Logsdail, "The Application of QM/MM Simulations in Heterogeneous Catalysis," *Physical Chemistry Chemical Physics* 25 (2023): 6562–6585.

25. J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters* 77 (1996): 3865.

26. F. Weigend and R. Ahlrichs, "Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy," *Physical Chemistry Chemical Physics* 7 (2005): 3297–3305.

27. F. Neese, "The SHARK Integral Generation and Digestion System," *Journal of Computational Chemistry* 44 (2023): 381–396.

28. P. P. Ewald, "Die Berechnung Optischer und Elektrostatischer Gitterpotentiale," *Annalen der Physik* 369 (1921): 253–287.

29. T. Darden, D. York, and L. Pedersen, "Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems," *Journal of Chemical Physics* 98 (1993): 10089–10092.

30. L. Greengard and V. Rokhlin, "A Fast Algorithm for Particle Simulations," *Journal of Computational Physics* 73 (1987): 325–348.

31. K. N. Kudin and G. E. Scuseria, "A Fast Multipole Method for Periodic Systems With Arbitrary Unit Cell Geometries," *Chemical Physics Letters* 283 (1998): 61–68.

32. M. Challacombe, C. White, and M. Head-Gordon, "Periodic Boundary Conditions and the Fast Multipole Method," *Journal of Chemical Physics* 107 (1997): 10131–10140.

33. K. N. Kudin and G. E. Scuseria, "Revisiting Infinite Lattice Sums With the Periodic Fast Multipole Method," *Journal of Chemical Physics* 121 (2004): 2886–2890.

34. M. Scheurer, P. Reinholdt, J. M. H. Olsen, et al., "Efficient Open-Source Implementations of Linear-Scaling Polarizable Embedding: Use Octrees to Save the Trees," *ChemRxiv* 17, no. 6 (2021): 3445–3454.

35. P. Reinholdt, J. Kongsted, and F. Lipparini, "Fast Approximate but Accurate QM/MM Interactions for Polarizable Embedding," *Journal of Chemical Theory and Computation* 18, no. 1 (2021): 344–356.

36. F. Lipparini, L. Lagardère, G. Scalmani, et al., "Quantum Calculations in Solution for Large to Very Large Molecules: A New Linear Scaling QM/Continuum Approach," *Journal of Physical Chemistry Letters* 5, no. 5 (2014): 953–958.

37. M. A. Watson, P. Sałek, P. Macak, and T. Helgaker, "Linear-Scaling Formation of Kohn-Sham Hamiltonian: Application to the Calculation of Excitation Energies and Polarizabilities of Large Molecular Systems," *Journal of Chemical Physics* 121, no. 7 (2004): 2915–2931.

38. D. Loco, É. Polack, S. Caprasecca, et al., "A QM/MM Approach Using the AMOEBA Polarizable Embedding: From Ground State Energies to Electronic Excitations," *Journal of Chemical Theory and Computation* 12, no. 8 (2016): 3654–3661.

39. F. Lipparini, G. Scalmani, L. Lagardère, et al., "Quantum, Classical, and Hybrid QM/MM Calculations in Solution: General Implementation of the ddCOSMO Linear Scaling Strategy," *Journal of Chemical Physics* 141, no. 18 (2014): 184108.

40. S. Caprasecca, S. Jurinovich, L. Lagardère, B. Stamm, and F. Lipparini, "Achieving Linear Scaling in Computational Cost for a Fully Polarizable MM/Continuum Embedding," *Journal of Chemical Theory and Computation* 11, no. 2 (2015): 694–704.

41. M. Nottoli, M. Bondanza, P. Mazzeo, et al., "QM/AMOEBA Description of Properties and Dynamics of Embedded Molecules," *Wiley Interdisciplinary Reviews: Computational Molecular Science* 13 (2023): e1674.

42. M. Bondanza, M. Nottoli, L. Cupellini, F. Lipparini, and B. Mennucci, "Polarizable Embedding QM/MM: The Future Gold Standard for Complex (Bio) Systems?," *Physical Chemistry Chemical Physics* 22 (2020): 14433–14448.

43. F. Lipparini, "General Linear Scaling Implementation of Polarizable Embedding Schemes," *Journal of Chemical Theory and Computation* 15 (2019): 4312–4317.

44. H. G. Petersen, D. Soelvason, J. W. Perram, and E. R. Smith, "The Very Fast Multipole Method," *Journal of Chemical Physics* 101 (1994): 8870–8876.

45. I. Gargantini, "An Effective Way to Represent Quadtree," *Communications of the ACM* 25 (1982): 905–910.

46. T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Hoboken, NJ: John Wiley & Sons, 2013).

47. C. A. White and M. Head-Gordon, "Derivation and Efficient Implementation of the Fast Multipole Method," *Journal of Chemical Physics* 101 (1994): 6593–6605.

48. F. Neese, "Software Update: The ORCA Program System—Version 5.0," *WIREs Computational Molecular Science* 12 (2022): e1606.

49. V. Anisimov and J. J. Stewart, *Introduction to the Fast Multipole Method: Topics in Computational Biophysics, Theory, and Implementation* (Boca Raton, FL: CRC Press, 2019).

50. J. M. Pérez-Jordá and W. Yang, "A Concise Redefinition of the Solid Spherical Harmonics and Its Use in Fast Multipole Methods," *Journal of Chemical Physics* 104 (1996): 8003–8006.

51. F. Neese, P. Colinet, B. DeSouza, B. Helmich-Paris, F. Wennmohs, and U. Becker, “The “Bubblepole” (BUPO) Method for Linear-Scaling Coulomb Matrix Construction With or Without Density Fitting,” *Journal of Physical Chemistry* (2024).

52. L. E. McMurchie and E. R. Davidson, “One- and Two-Electron Integrals Over Cartesian Gaussian Functions,” *Journal of Computational Physics* 26 (1978): 218–231.

53. C. A. White and M. Head-Gordon, “Rotating Around the Quartic Angular Momentum Barrier in Fast Multipole Method Calculations,” *Journal of Chemical Physics* 105 (1996): 5061–5067.

54. C. Lee, W. Yang, and R. G. Parr, “Development of the Colle-Salvetti Correlation-Energy Formula Into a Functional of the Electron Density,” *Physical Review B* 37 (1988): 785.

55. A. D. Becke, “A New Mixing of Hartree–Fock and Local Density-Functional Theories,” *Journal of Chemical Physics* 98 (1993): 1372–1377.

56. J. P. Perdew, J. Chevary, S. Vosko, et al., “Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation,” *Physical Review B* 48 (1993): 4978.

57. D. Goldberg, “What Every Computer Scientist Should Know About Floating-Point Arithmetic,” *ACM Computing Surveys* 23 (1991): 5–48.

58. S. Bhattacharjee, F. Neese, and D. A. Pantazis, “Triplet States in the Reaction Center of Photosystem II,” *Chemical Science* 14 (2023): 9503–9516.

59. A. Sirohiwal, F. Neese, and D. A. Pantazis, “Protein Matrix Control of Reaction Center Excitation in Photosystem II,” *Journal of the American Chemical Society* 142 (2020): 18174–18190.

60. A. Sirohiwal and D. A. Pantazis, “The Electronic Origin of Far-Red-Light-Driven Oxygenic Photosynthesis,” *Angewandte Chemie, International Edition* 61 (2022): e202200356.

61. J.-D. Chai and M. Head-Gordon, “Long-Range Corrected Hybrid Density Functionals With Damped Atom–Atom Dispersion Corrections,” *Physical Chemistry Chemical Physics* 10 (2008): 6615.

62. S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, “A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu,” *Journal of Chemical Physics* 132 (2010): 154104.

63. F. Weigend, F. Furche, and R. Ahlrichs, “Gaussian Basis Sets of Quadruple Zeta Valence Quality for Atoms H-Kr,” *Journal of Chemical Physics* 119 (2003): 12753–12762.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.