Grassmann extrapolation of density matrices for Born-Oppenheimer molecular dynamics

Etienne Polack, Geneviève Dusson, Benjamin Stamm, Filippo Lipparini

Grassmann extrapolation of density matrices for Born-Oppenheimer molecular dynamics

Étienne Polack,[†] Geneviève Dusson,[†] Benjamin Stamm,[‡] and Filippo Lipparini[*,¶]

[†]*Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université Bourgogne Franche-Comté, 16 route de Gray, 25030 Besançon, France*

[‡]*Department of Mathematics, RWTH Aachen University, Schinkelstr. 2, 52062 Aachen, Germany*

[¶]*Dipartimento di Chimica e Chimica Industriale, Univeristà di Pisa, Via G. Moruzzi 13, I-56124 Pisa, Italy*

E-mail: filippo.lipparini@unipi.it

Abstract

Born–Oppenheimer Molecular Dynamics (BOMD) is a powerful but expensive technique. The main bottleneck in a density functional theory BOMD calculation is the solution to the Kohn–Sham (KS) equations, that requires an iterative procedure that starts from a guess for the density matrix. Converged densities from previous points in the trajectory can be used to extrapolate a new guess, however, the non-linear constraint that an idempotent density needs to satisfy make the direct use of standard linear extrapolation techniques not possible. In this contribution, we introduce a locally bijective map between the manifold where the density is defined and its tangent space, so that linear extrapolation can be performed in a vector space while, at the same time, retaining the correct physical properties of the extrapolated density using molecular descriptors. We apply the method to real-life, multiscale polarizable QM/MM

BOMD simulations, showing that sizeable performance gains can be achieved, especially when a tighter convergence to the KS equations is required.

1 Introduction

Ab-initio Born–Oppenheimer molecular dynamics (BOMD) is one of the most powerful and versatile techniques in computational chemistry, but its computational cost represents a big limitation to its routine use in quantum chemistry. To perform a BOMD simulation, one needs to solve the quantum mechanics (QM) equations, usually Kohn–Sham (KS) density functional theory (DFT), at each step, before computing the forces and propagating the trajectory of the nuclei. The iterative self-consistent field (SCF) procedure is expensive, as it requires to build at each iteration the KS matrix and to diagonalize it. Convergence can require tens of iterations, making the overall procedure, which has to be repeated a very large number of times, very expensive. To reduce the cost of BOMD simulations, it is therefore paramount to be able to perform as little iterations as possible while, at the same time, obtaining an SCF solution accurate enough to afford stable dynamics.

From a conceptual point of view, at each step of a BOMD simulation, a map is built from the molecular geometry to the SCF density, and then to the energy and forces. The former map, in practice, requires the solution to the SCF problem and is not only very complex, but also highly non-linear. However, the propagation of the molecular dynamics (MD) trajectory uses short, finite time steps, so that the converged densities at previous steps, and thus at similar geometries, are available. As a consequence, the geometry to density map can be in principle approximated by extrapolating the available densities at previous steps. The formulation of effective extrapolation schemes has been the object of several previous works.[1] Among the proposed strategies, one for density matrix extrapolation was developed by Alfè[2], as a generalization of the wavefunction extrapolation method by Arias et al.[3], which is based on a least-squares regression on a few previous atomic positions. The main difficulty

in performing an extrapolation of the density matrix stems from the non-linearity of the problem. In other words, a linear combination of idempotent density matrices is not an idempotent density matrix, as density matrices are elements of a manifold and not of a vector space. To circumvent this problem, strategies that extrapolate the Fock or KS matrix[4,5] or that use orbital transformation methods[6–8] have been proposed.

A completely different strategy has been proposed by Niklasson and coworkers.[9–11] In the extended Lagrangian Born–Oppenheimer (XLBO) method, an auxiliary density is propagated in a time-reversible fashion and then used as a guess for the SCF procedure. The strategy is particularly successful, as it combines an accurate guess with excellent stability properties. In particular, the XLBO method allows one to perform accurate simulations converging the SCF to average values (for instance, 10^{-5} in the root-mean-square (RMS) norm of the density increment), which are usually insufficient to compute accurate forces. An XLBO-based BOMD strategy has been recently developed by some of us in the context of polarizable multiscale BOMD simulations of both ground and excited states.[12–15] Multiscale strategies can be efficiently combined, in a focused model spirit, to BOMD simulations to extend the size of treatable systems. Using a polarizable embedding allows one to achieve good accuracy in the description of environmental effects, especially if excited states or molecular properties are to be computed. In such a context, the XLBO guessing strategy allows one to perform stable simulation even using the modes 10^{-5} RMS convergence threshold, which, thanks to the quality of the XLBO guess, typically requires only about 4 SCF iterations. Recently, SCF-less formulations of the XLBO schemes have also been proposed.[16,17]

Unfortunately, the performances of the XLBO-based BOMD scheme are not so good when a tighter SCF convergence is required, which can be the case when one wants to perform MD simulations using post Hartree–Fock (HF) methods or for excited states described in a time-dependent DFT framework.[14,18] In fact, such methods require the solution to a second set of QM equations which are typically non-variational, making them more susceptible to numerical errors and instabilities. Computing the forces for non-SCF energies requires therefore a more

accurate SCF solution.

The present work builds on all previous methods for density matrix extrapolation and aims at proposing a simple framework to overcome the difficulties associated with the non-linearity of the problem. The strategy that we propose is based on a differential geometry approach and is particularly simple. First, we introduce a molecular descriptor, i.e., a function of the molecular geometry and other molecular parameters that represents the molecular structure in a natural way that respects the invariance properties of the molecule within a vector space. At the $(n+1)$-th step of an MD trajectory, we fit the new descriptor in a least-square fashion using the descriptors available at a number of previous steps and obtain a new set of coefficients. However, we do not use them to directly extrapolate the density. Instead, we first map the unknown density matrix, that we aim to approximate, from the manifold where it is defined to its tangent space. We then perform the extrapolation to approximate the representative density matrix in the tangent space, before mapping this approximation back to the manifold in order to obtain an extrapolated density matrix that satisfies the required physical constraints. This geometrical strategy, that has recently been introduced in the context of density matrix approximation by us,[19] allows one to use standard linear extrapolation machinery without worrying about the non-linear physical constraints on the density matrix, since both the space of descriptors and the tangent space are vector spaces. As the mapping between the manifold and the tangent space is locally bijective, no concerns about redundant degrees of freedom (such as rotations that mix occupied orbitals) arise. The map and its inverse, which are known as Grassmann Logarithm and Exponential, are easily computed and the implementation of the strategy is straightforward. We shall denote this approach as Grassmann extrapolation (G-Ext).

In this contribution, we choose a simple, yet effective molecular descriptor and, for the extrapolation, a least square strategy. These are not the only choices. As our strategy allows one to use any linear extrapolation technique between two vector spaces, which can be in turn coupled with any choice of molecular descriptor, more advanced strategies can

be proposed, including machine learning. Our approach ensures that the extrapolated density, independent of how it is obtained, satisfies all the physical requirements of a density stemming from a single Slater determinant.

The paper is organized as follows. In the upcoming Section 2, we present all necessary theoretical foundations required for the development and implementation of the presented G-Ext approach. Section 3 then presents detailed numerical tests illustrating the performance of the extrapolation scheme, including realistic applications of BOMD within a QM/molecular mechanics (MM)-context before we draw the conclusion in Section 4.

2 Theory

We consider Born–Oppenheimer ab-initio BOMD simulations where the position vector $\mathsf{R} \in \mathbb{R}^{3M}$ evolves in time according to classical mechanics as

$$M_i \ddot{\mathsf{R}}_i(t) = F_i(t, \mathsf{R}(t)), \tag{1}$$

where $\mathsf{R}_i(t), F_i(t) \in \mathbb{R}^3$ denote the position of the i-th atom with mass M_i respectively the force acting on it at time t. We consider a general QM/MM-method but the setting also trivially applies to pure QM-models. The forces at a given time t and position R of the nuclei arise from different interactions, namely QM-QM, QM-MM and MM-MM interactions. The computationally expensive part is to determine the state of the electronic structure, which is modelled here at the DFT level with a given basis set of dimension $\mathcal{N}$. Note that considering HF instead of DFT would not change much in the presentation of the method. It consists of computing the instantaneous non-linear eigenvalue problem

$$\begin{cases} F_{\mathsf{R}}(D_{\mathsf{R}})C_{\mathsf{R}} = S_{\mathsf{R}}C_{\mathsf{R}}E_{\mathsf{R}} \\ C_{\mathsf{R}}^{\mathsf{T}} S_{\mathsf{R}} C_{\mathsf{R}} = \mathrm{Id}_N \\ D_{\mathsf{R}} = C_{\mathsf{R}}C_{\mathsf{R}}^{\mathsf{T}} \end{cases}, \tag{2}$$

where $C_{\mathsf{R}} \in \mathbb{R}^{\mathcal{N}\times N}$ and $D_{\mathsf{R}} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ denote the coefficients respectively of the occupied orbitals and density matrix and $E_{\mathsf{R}} \in \mathbb{R}^{\mathcal{N}\times N}$ the diagonal matrix containing the energy levels. Further, F_{R} denotes the DFT-operator acting on the density matrix and S_{R} the customary overlap matrix.

At this point it is useful to note that the slightly modified coefficient matrix $\widetilde{C}_{\mathsf{R}} \coloneqq S_{\mathsf{R}}^{1/2} C_{\mathsf{R}}$ belongs to the so-called Stiefel manifold defined as follows

$$\mathcal{S}t(N,\mathcal{N}) \coloneqq \left\{ V \in \mathbb{R}^{\mathcal{N}\times N} \;\middle|\; V^{\mathsf{T}} V = \mathrm{Id}_N \right\}, \tag{3}$$

due to the second equation in Equation (2). In consequence the normalized density matrix $\widetilde{D}_{\mathsf{R}} = \widetilde{C}_{\mathsf{R}} \widetilde{C}_{\mathsf{R}}^{\mathsf{T}} = S_{\mathsf{R}}^{1/2} D_{\mathsf{R}} S_{\mathsf{R}}^{1/2}$ belongs to the following set

$$\mathcal{G}r(N,\mathcal{N}) \coloneqq \left\{ D \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}} \;\middle|\; D^2 = D,\ D^{\mathsf{T}} = D,\ \mathrm{Tr}\, D = N \right\}, \tag{4}$$

which can be identified with the Grassmann manifold of N-dimensional subspaces of $\mathbb{R}^{\mathcal{N}}$ by means of the spectral projectors. In the following, we always assume that the density matrix has been orthonormalized, and therefore drop the $\sim$ from the notation. For any $D \in \mathcal{G}r(N,\mathcal{N})$, one can associate the tangent space $\mathcal{T}_D$ which has the structure of a vector space. The evolution of the electronic structure can therefore be seen as a trajectory $t \mapsto D_{\mathsf{R}(t)}$ on $\mathcal{G}r(N,\mathcal{N})$ where $t \mapsto \mathsf{R}(t)$ denotes the trajectory of the nuclei.

The goal of the present work is to find a good approximation for the electronic density matrix at the next step of MD trajectory by extrapolating the densities at previous steps. More precisely, based on the knowledge of the density matrices $D_i \coloneqq D_{\mathsf{R}(t_i)}$, $i = n - N_t, \ldots, n-1$, at N_t previous times t_i, one aims to compute an accurate guess of the density matrix D_n at time t_n.

Thus, the problem formulation can be seen as an extrapolation problem of the following form: given the set of couples $(\mathsf{R}(t_i), D_i)$ and a new position vector $\mathsf{R}(t_n)$, provide a guess for the solution D_n. Here and in the remaining part of the article, we restrict ourselves on

the positions of the QM-atoms, i.e., with slight abuse of notation we denote from now on by R the set of QM-positions only, even within a QM/MM-context.

In order to approximate the mapping $\mathsf{R} \mapsto D_{\mathsf{R}}$, we split this mapping in several sub-maps that will be composed as follows:

$$\begin{aligned} \mathbb{R}^{3M} \to \mathcal{M} \to \mathcal{T}_{D_0} \to \mathcal{G}r(N, \mathcal{N}) \\ \mathsf{R} \mapsto d_{\mathsf{R}} \mapsto \Gamma_{\mathsf{R}} \mapsto D_{\mathsf{R}} = \mathrm{Exp}_{D_0}(\Gamma_{\mathsf{R}}), \end{aligned} \tag{5}$$

where the first line shows the concatenation of maps in terms of spaces and the second in terms of variables. The different mappings will be presented and motivated in the following.

The first map is a mapping of the nuclear coordinates $\mathsf{R} \in \mathbb{R}^{3M}$ to a (possibly high-dimensional) molecular descriptor $d_{\mathsf{R}} \in \mathcal{M}$ that accounts for certain symmetries and invariances of the molecule. The last map, known as the Grassmann exponential, is introduced in order to obtain a resulting density matrix belonging to $\mathcal{G}r(N, \mathcal{N})$ and thus to guarantee that the guess fulfils all properties of a density matrix. As $\mathcal{G}r(N, \mathcal{N})$ is a manifold this is not straightforward. The second mapping is the one that we aim to approximate but before we do that, let us first introduce those two special mappings, i.e., the molecular descriptor and the Grassmann exponential, in more details.

2.1 Molecular descriptors

The map $\mathsf{R} \mapsto d_{\mathsf{R}}$ is a map from atomic positions to molecular descriptors. These descriptors are used as fingerprints for the considered molecular configurations. Such molecular descriptors have been widely used in the past decades e.g., to learn potential energy surfaces (PES),[20–26] or to predict other quantities of interest. Among widely used descriptors, one can find Behler–Parinello symmetry functions,[27] Coulomb matrix,[28] smooth overlap of atomic positions (SOAP),[29] permutationally invariant polynomials,[30] or the atomic cluster expansion (ACE).[31,32] These molecular descriptors are usually designed to retain similar symmetries as the targeted quantities of interest.

In this work, the quantity we are approximating is the density matrix, which is invariant with respect to translations as well as permutations of like particles. The transformation of the density matrix with respect to a global rotation of the system depends on the implementation, as it is possible to consider either a fixed Cartesian frame or one that moves with respect to the molecular system. In the former case, there is an equivariance with respect to rotations of the molecular system, while in the latter, the density matrix is invariant. We should therefore in principle use a molecular descriptor satisfying those properties.

However, the symmetry properties we will rely on are mostly translation and rotation invariance. Therefore, we will use a simple descriptor in form of the Coulomb-matrix denoted by d_{R}, given by

$$(d_{\mathsf{R}})_{ij} = \begin{cases} 0.5 z_i^{2.4} & \text{if } i = j \\ \dfrac{z_i z_j}{\|\mathsf{R}(t_i) - \mathsf{R}(t_j)\|} & \text{otherwise} \end{cases} . \tag{6}$$

Note that such a descriptor is not invariant (nor equivariant) with respect to permutations of identical particles. However, we have found this descriptor to offer a good trade-off between simplicity and efficiency. Note that since we aim to extrapolate the density matrix from previous time-steps, permutations of identical particles never occur from one time-step to another and we do not need to rely on this property. Nevertheless, we expect that a better description could be achieved by using more flexible descriptors, such as ACE polynomials or the SOAP descriptors, where the descriptors themselves can be tuned.

2.2 The Grassmann exponential

We only give a brief overview as the technical details have already been reported elsewhere.[19,33,34] The set $\mathcal{G}r(N,\mathcal{N})$ is a smooth manifold and thus, at any point, say $D_0 \in \mathcal{G}r(N,\mathcal{N})$ in our application, there exists the tangent space $\mathcal{T}_{D_0}$ such that one can associate nearby points $D \in \mathcal{G}r(N,\mathcal{N})$ to tangent vectors $\Gamma(D) \in \mathcal{T}_{D_0}$. The mapping $D \mapsto \mathrm{Log}_{D_0}(D) = \Gamma(D)$ is known as the Grassmann logarithm and its inverse mapping as the Grassmann exponential $\Gamma \mapsto \mathrm{Exp}_{D_0}(\Gamma) = D$. There also holds that $\mathrm{Log}_{D_0}(D_0) = 0$ and

$\mathrm{Exp}_{D_0}(0) = D_0$. These mappings are not only abstract tools from differential geometry but can be computed by means of performing an singular value decomposition (SVD).[19,33,34] In our application we use the same reference point D_0 in all cases which brings some computational advantages as will be discussed in more detail the the upcoming Section 2.3.

2.3 The approximation problem

Since the tangent space $\mathcal{T}_{D_0}$ is a (linear) vector space, we can now aim to approximate the mapped density matrix on the tangent space $\mathcal{T}_{D_0}$. To simplify the presentation, we shift the indices in the following and describe the extrapolation method for the first N_t time steps. In the general setting, we should consider the positions $\mathsf{R}(t_i)$ for $i = n - N_t, \ldots, n-1$, to extrapolate the density matrix at position $\mathsf{R}(t_n)$, where n is the current time step of the MD. We look for parameter functions c_i, such that, given previous snapshots $\Gamma_i = \mathrm{Log}_{D_0}(D_i)$ for i from 1 to N_t, corresponding to some $\mathsf{R}(t_i)$'s, the approximation of any density matrix on the tangent space is written as

$$\mathsf{R} \mapsto \Gamma_{\mathrm{app}}(\mathsf{R}) = \sum_{i=1}^{N_t} c_{\mathsf{R},i}\, \Gamma_i \in \mathcal{T}_{D_0}, \tag{7}$$

with $\Gamma_i = \Gamma_{\mathsf{R}(t_i)}$.

The question is then how to find these coefficient functions $c_{\mathsf{R},i}$ and we propose to find those via the resolution of a (standard) least-square minimization problem. For a given position R, we look for coefficients that minimise the ℓ^2–error between the descriptor d_{R} and a linear combination of the previous ones $d_{\mathsf{R}(t_i)}$

$$\min_{c_{\mathsf{R}} \in \mathbb{R}^{N_t}} \left\| d_{\mathsf{R}} - \sum_{i=1}^{N_t} c_{\mathsf{R},i} d_{\mathsf{R}(t_i)} \right\|^2. \tag{8}$$

In matrix form, this simply reads

$$\min_{c_{\mathsf{R}} \in \mathbb{R}^{N_t}} \quad \left\| d_{\mathsf{R}} - P^{\mathsf{T}} c_{\mathsf{R}} \right\|^2, \tag{9}$$

where P is the matrix of size $N_t \times N_d$ containing the descriptors $P_{i,j} \coloneqq (d_{\mathsf{R}(t_i)})_j$. Note that we only fit on the level of the descriptor, i.e., the mapping from the position vector R to the descriptor d_{R}, and that this method is similar to the ones used by Alfè [2], Arias et al. [3], where the descriptors they used were the positions of the atoms and only considered the previous three time-steps of the MD.

If the system is underdetermined, we select the vector c_{R} that has the smallest norm. However, in general, the system is overdetermined as we have more descriptors than snapshots. This implies that this formulation verifies the interpolation principle: for every i and j from 1 to N_t, the solution of Problem (8) at the positions $\mathsf{R}(t_j)$ satisfies $c_{\mathsf{R}(t_j),i} = \delta_{ji}$.

In principle, should we consider a large amount of previous descriptors, then the system may become undetermined and violates the interpolation principle. To mitigate this, we can use a stabilization scheme, as explained in the upcoming subsection.

Note that once we have computed the coefficients c_{R} by solving Problem (12), one computes the initial guess for the density by using the same coefficients in the linear combination on the tangent space as in Equation (7) and finally take the exponential (see Equation (5)). The rational for this step is that, if the second mapping in Equation (5), that we denote here by $\mathcal{F}\colon \mathcal{M} \to \mathcal{T}_{D_0}$, was linear, then there would hold

$$\mathcal{F}\left(\sum_{i=1}^{N_t} c_{\mathsf{R},i} d_{\mathsf{R}_i} \right) = \sum_{i=1}^{N_t} c_{\mathsf{R},i} \mathcal{F}(d_{\mathsf{R}_i}) = \sum_{i=1}^{N_t} c_{\mathsf{R},i} \Gamma_i. \tag{10}$$

In practice, the mapping is however not linear and this approach works well in the test cases we considered. A possible explanation for this is the unfolding of the nuclear coordinates into a high-dimensional descriptor-space $\mathcal{M}$. Indeed, the high-dimensionality of $\mathcal{M}$ seems to allow an accurate approximation of $\mathcal{F}$ by a linear map. Further, if the system is overdetermined,

the scheme satisfies the interpolation property $\Gamma_j = \Gamma(\mathsf{R}(t_j))$, and hence we recover the expected density matrix $D_{\mathsf{R}(t_j)} = \mathrm{Exp}_{D_0}(\Gamma_j)$.

2.3.1 Stabilization

To stabilize the extrapolation by limiting high oscillations of the coefficients, we apply a Tikhonov regularization

$$\min_{c_{\mathsf{R}} \in \mathbb{R}^{N_t}} \left(\left\| d_{\mathsf{R}} - \sum_{i=1}^{N_t} c_{\mathsf{R},i} d_{\mathsf{R}_i} \right\|^2 + \varepsilon \left\| c_{\mathsf{R}} \right\|^2 \right), \tag{11}$$

for some choice of ε. This problem is always well-posed, and corresponds to solving the following problem

$$\min_{c_{\mathsf{R}} \in \mathbb{R}^{N_t}} \left\| \widetilde{d_{\mathsf{R}}} - \widetilde{P}^{\mathsf{T}} \cdot c_{\mathsf{R}} \right\|^2, \tag{12}$$

where $\widetilde{d_{\mathsf{R}}} \in \mathbb{R}^{N_d+N_t}$ is the vector d_{R} padded with N_t zeros and $\widetilde{P} \in \mathbb{R}^{N_t} \times \mathbb{R}^{N_d+N_t}$ is the P matrix padded with the square diagonal matrix $\varepsilon \, \mathrm{Id}_{N_t}$. We observe in practice that using such a stabilization makes possible to use more previous points without degradation of the initial guess.

2.4 The final algorithm

Given previous density matrices $D_{\mathsf{R}(t_j)}$ for $j = 1, \ldots, N_t$, the initial guess is computed following Algorithm 1. That is, we start by computing the logarithms of the density matrices $D_{\mathsf{R}(t_j)}$, from the coefficients $C_{\mathsf{R}(t_j)}$ that are first orthonormalized by performing $\widetilde{C}_{\mathsf{R}} = S_{\mathsf{R}}^{1/2} C_{\mathsf{R}}$. Here, we remark that we assume that the density matrices $D_{\mathsf{R}(t_j)}$ have been previously Löwdin orthonormalized.

We then compute the descriptors needed to build the $\widetilde{P}$ matrix and solve Problem (12). This provides the coefficients in the linear combination of the $\Gamma_i's$ on the tangent space. Finally, we compute the exponential of the linear combination in order to obtain the predicted

Algorithm 1: Density extrapolation framework G-Ext

Data: Array desc containing the descriptors for k previous time-steps, p_n the descriptor for the current position, C_{n-1} and S_{n-1} respectively the molecular orbitals and overlap matrices of the previous time-step, and cref the reference point on the Grassmannian

Result: Guess density matrix for time-step $n > 1$

begin

$\mathsf{cmat}(:, :, n-1) \leftarrow \mathtt{Orthonormalization}(C_{n-1}, S_{n-1})$;
$\mathsf{gmat}(:, :, n-1) \leftarrow \mathtt{Log}(\mathsf{cref}, \mathsf{cmat}(:, :, n-1))$;
$\mathsf{desc}, p_n \leftarrow \mathtt{Stabilization}(\mathsf{desc}, p_n)$;
$\mathsf{c} \leftarrow \mathtt{LeastSquares}(\mathsf{desc}, p_n)$;
$\Gamma_{\text{app}} \leftarrow \sum_{i=n-1-k}^{n-1} \mathsf{c}(i) \cdot \mathsf{gmat}(:, :, i)$;
$C_{\text{app}} \leftarrow \mathtt{Exp}(\mathsf{cref}, \Gamma_{\text{app}})$;
return $2 \cdot C_{\text{app}} \cdot C_{\text{app}}^{\mathsf{T}}$;

density matrix.

Note that the reference point D_0 is chosen once and for all, which makes the computations of these logarithms lighter, even though there is no theoretical justification for keeping a single point D_0 as a reference. Indeed, it is known that the formulae are only correct locally (around D_0) on the manifold. However, in practice we have never observed the need to change the reference point. This enables us to compute only one logarithmic map per time step; and hence, only two SVD in total per time step. To have a robust algorithm that will work even in this edge case, it will be sufficient to check that the exponential and logarithmic maps are still inverse of one another.

Finally, to be on the safe-side with respect to the computations of the exponential, we have added a check on the orthogonality of the matrix that is obtained: If the residue is higher than a certain threshold, we then perform an orthogonalization of the result.

3 Numerical tests

In this section we present a series of numerical tests of the newly developed strategy. We test our method on four different systems. All the systems have been object of a previous

or current study by some of us, and can therefore be considered representative of real-life applications.

The first system is 3-hydroxyflavone (3HF) in acetonitrile.[18] Two systems (OCP and APPA) are chromophores embedded in a biological matrix — namely, a carotenoid in the orange carotenoid protein (OCP) and flavine in acid phosphatase (APPA), a blue light-using flavine photoreceptor.[35–37] The fourth system is dimethylaminobenzonitrile (DMABN) in methanol.[14] The main characteristics of the systems used for testing are recapitulated in Table 1.

Table 1: Overview of the system size in terms of number of QM-atoms (N_{QM}), number of MM-atoms (N_{MM}) and the total number of (QM) basis functions ($\mathcal{N}$).

System	N_{QM}	N_{MM}	$\mathcal{N}$
OCP	129	4915	1038
APPA	31	16 449	309
DMABN	21	6843	185
3HF	28	15 018	290

The systems used for testing include a quite large QM chromophore, the OCP and three medium-sized systems, embedded in large (APPA, 3HF) and medium-sized environments (DMABN) and are representative of different possible scenarios.

To test the performances of the new G-Ext strategy, we performed three sets of short (1 ps) multiscale BOMD simulations on OCP, APPA, 3HF, and DMABN. KS density functional theory was used to model the QM subsystem, using the B3LYP[38] hybrid functional and Pople's 6-31G(d) basis set.[39] For the stability and energy conservation of the method, we did a longer and more realistic simulations of 10 ps on 3HF, where the flavone moiety was described using the ωB97X hybrid functional[40] and Pople's 6-31G(d) basis set. In all cases, the environment was modeled using the AMOEBA polarizable force field.[41]

All the simulations have been performed using the Gaussian–Tinker interface previously developed by some of us.[12,13] In particular, we use a locally modified development version of Gaussian[42] to compute the QM, electrostatic and polarization energy and forces, and Tinker[43] to compute all others contributions to the QM/MM energy. We implemented the G-Ext

extrapolation scheme in Tinker, that acts as the main driver for the MD simulation, being responsible of summing together all the various contributions to the forces and propagating the trajectory. At each MD step, using the GauOpen interface,[44] the density matrix, molecular orbital (MO) coefficients, and overlap matrix produced by Gaussian are retrieved. These are used to compute the extrapolated density as described in Section 2. The density is then passed back to Gaussian to be used for the next MD step. All the simulations were carried out in the NVE ensemble, using the velocity Verlet integrator and a 0.5 fs time step. Concerning stabilization, we found that good overall results were obtained using a parameter $\varepsilon := 10^3 \cdot r_{\text{scf}}$, where r_{scf} is the tolerance of the SCF algorithm.

3.1 Numerical results

To assess the performance of the G-Ext guess we perform 1 ps MD simulations on the four systems described in Section 3 starting from the same exact conditions (positions and initial velocities) and using various strategies to compute the guess density for the SCF solver. We compare various flavors of the G-Ext method with the the XLBO extrapolation scheme.[10] Here, we note that the original XLBO method performs a propagation of an auxiliary density matrix, which is then used as a guess. The latter is not idempotent: to restore such a property, we perform a purification step at the beginning of the SCF procedure using McWeeny's algorithm.[45] In the following, we therefore compare our method, where we use 3 to 6 previous points for the fitting and extrapolation, to both the standard XLBO and to XLBO followed by purification (XLBO/MW). We use an SCF convergence threshold of 10^{-5} with respect to the RMS variation of the density.

We report in Table 2, for each method, the average number of SCF iterations performed along the MD simulation together with the associated standard deviation. As the XLBO strategy requires 8 previous points, during which a standard SCF is performed, we discard the first points from the evaluation of the aforementioned quantities to have a fairer comparison.

We do not report the total time required to compute the guess, as it is in all cases very

small (up to 0.1 s wall clock time for the largest system using the G-Ext(6) guess). This is an important consideration, as the G-Ext method requires one to perform various linear-algebra operations (in particular, thin SVD) that can in principle be expensive. Thanks to the availability of optimized LAPACK libraries, this is in practice not a problem.

Table 2: Performances of the G-Ext method for different number of extrapolation points, compared with the XLBO algorithm with and without McWeeny purification. All the results were obtained using a 10^{-5} convergence threshold for the root-mean-square increment of the density matrix and are derived from a 1 ps long MD simulation, using a 0.5 fs time step. We report the average number of iterations required to converge the SCF, together with the associated standard deviation. Note that the first 8 steps were discarded.

	OCP		DMABN		APPA		3HF	
Method	Average	σ	Average	σ	Average	σ	Average	σ
XLBO	3.82	0.66	3.98	0.16	3.00	0.03	4.00	0.14
XLBO/MW	2.95	0.31	3.76	0.56	3.00	0.34	3.96	0.31
G-Ext(3)	2.57	0.84	3.54	0.78	2.95	0.50	3.09	0.41
G-Ext(4)	2.48	0.88	3.14	0.62	2.51	0.50	3.25	0.68
G-Ext(5)	2.25	0.96	3.23	0.75	2.51	0.50	3.30	0.72
G-Ext(6)	2.20	0.96	2.99	0.02	2.51	0.50	3.14	0.56

From the results in Table 2, we see that the G-Ext algorithms systematically outperforms the XLBO method. It is interesting to note that the McWeeny purification step has a sizeable effect on the performances of the XLBO method only for the largest system, OCP, where it results in the gain of almost one SCF iteration on average. On the other systems, the purification step has a smaller effect.

In all the systems we tested, the performances of the G-Ext method are systematically better than in XLBO, including with McWeeny purification. The effectiveness of the G-Ext extrapolation increases when going from 3 to 6 points, but quickly stagnates. We have performed further tests with more than 6 (up to 20) extrapolation points, but never noted any further gain.

We observe a reduction in the number of iterations that goes from 0.5 in DMABN to 0.75 in OCP (1.62 when compared to XLBO without McWeeny purification). We remark that these gains, while apparently not so large, are greatly amplified during the MD simulation, due to

the large number of steps that need to be performed.

The tests performed with a 10^{-5} convergence threshold are representative of a standard, DFT ground state BOMD simulation. When performing a more sophisticated quantum mechanical calculation, such as a BOMD on an excited state PES,[18] such a convergence threshold may not be sufficient to guarantee the stability of the simulation, as the SCF solution is used to set up the linear response equations and the numerical error can be amplified, resulting in poorly accurate forces.

We tested the G-Ext algorithm in its best-performing version, the one that uses six extrapolation points, with a tighter, 10^{-7} threshold, again for the RMS variation of the density. The results are reported in Table 3, where we compare the G-Ext(6) scheme with the XLBO method with McWeeny purification.

The XLBO method is based on the propagation of an auxiliary density and therefore the accuracy of the guess it generates depends little on the accuracy of the previous SCF densities. As a consequence, its performances are reduced if a tighter convergence is required. The G-Ext guess, on the other hand, uses previously computed densities as its building blocks and one can expect the accuracy of the resulting guess to be linked to the convergence threshold used during the simulation.

This is exactly what we observe. Using a threshold of 10^{-7}, the G-Ext(6) guess exhibits significantly better performances than XLBO, gaining, on average, from about 0.7 to about 3 SCF iterations on the tested systems.

3.1.1 Stability

The good performances of the G-Ext guess come, however, at a price, namely, the lack of time reversibility. We can thus expect the total energy in a NVE simulation to exhibit a long-time drift (LTD). Time reversibility and long-time energy conservation are, on the other hand, one of the biggest strengths of the XLBO method.

To investigate the stability of BOMD simulations using the G-Ext guess, we build a chal-

Table 3: Performances of the G-Ext(6) method compared with the XLBO algorithm with McWeeny purification. All the results were obtained using a 10^{-7} convergence threshold for the root-mean-square increment of the density matrix and are derived from a 1 ps long MD simulation, using a 0.5 fs time step. We report the average number of iterations required to converge the SCF, together with the associated standard deviation. Note that the first 8 steps were discarded.

	OCP		DMABN		APPA		3HF	
Method	Average	σ	Average	σ	Average	σ	Average	σ
XLBO/MW	5.02	0.17	7.30	0.64	7.49	0.84	7.47	0.63
G-Ext(6)	3.58	0.79	4.23	0.50	4.39	0.57	6.81	0.78

lenging case, where we start a BOMD simulation far from well-equilibrated conditions. We use the 3HF system as a test case and achieve the noisy starting conditions by starting from a well-equilibrated structure and changing the DFT functional from B3LYP to ωB97XD. This way, we have a physically acceptable structure, with no close atoms or other problematic structural situations, but obtain starting conditions that are far from equilibrium.

We report in Figure 1 the total energy along a 10 ps BOMD simulation of 3HF in acetonitrile using either a 10^{-5} SCF convergence threshold (left panel) or a 10^{-7} one (right panel). The same results for a 10^{-6} threshold are reported in the supporting information. We compare the G-Ext(3) and G-Ext(6) methods to the XLBO one including McWeeny purification. As already noted, while in principle the purification may spoil the time reversibility, this has no noticeable effect in practice.

The very noisy starting conditions are apparent from the energy profiles, that exhibits large oscillations in the first couple hundreds femtoseconds.

To better estimate the short- and long-time energy stability, we report in Table 4 the average short-time fluctuation (STF) and LTD of the energy. The former is computed by taking the RMS of the energy fluctuation every 50 fs and averaging the results over the trajectory, discarding the first 500 fs, the latter by fitting the energy with a linear function and taking the slope.

All methods show comparable short-term stability, which is to be mainly ascribed to

the chosen integration time-step. On the other hand, from both the results in Table 4 and Figure 1, we observe a clear drift of the energy when the G-Ext method is used. In particular, the system cools of about 10 kcal/mol with either G-Ext(3) or G-Ext(6). The XLBO trajectory, despite the McWeeny purification, exhibits an almost perfect energy conservation.

These results are not surprising, but should be taken into account when choosing to use the G-Ext guess, which, if coupled to a 10^{-5} SCF convergence threshold, cannot guarantee long-term energy conservation. The drift is overall not too large and can be handled by using a thermostat. Whether or not the trade between performances and energy conservation is acceptable for a production simulation is a decision that ultimately lies with the user.

Increasing the accuracy of the SCF computation improves the overall stability for G-Ext, which is already good at 10^{-6} and becomes virtually identical to the one offered by the XLBO method at 10^{-7}.

Table 4: Short and long-term stability analysis of the G-Ext(3) and G-Ext(6) methods, compared to the XLBO algorithm with McWeeny purification, for the 3HF system. For each method we report the STF and the LTD and the average number of SCF iterations, for three convergence thresholds of the SCF algorithm.

	Conv. 10^{-5}		Conv. 10^{-6}		Conv. 10^{-7}	
Method	STF	LTD	STF	LTD	STF	LTD
XLBO/MW	0.55	−0.04	0.55	−0.03	0.57	−0.03
G-Ext(3)	0.55	−0.42	0.57	−0.15	0.53	−0.04
G-Ext(6)	0.56	−0.53	0.52	−0.13	0.57	−0.04

4 Conclusion

In this contribution, we presented an extrapolation scheme to predict initial guesses of the density matrix for the SCF-iterations within BOMD. What makes our approach new is that we enforce the idempotency of the density matrix by extrapolating not the densities themselves, but their map onto a vector space, which is the tangent plane to the manifold of the physically acceptable densities. Such a map is locally bijective, so that after performing the

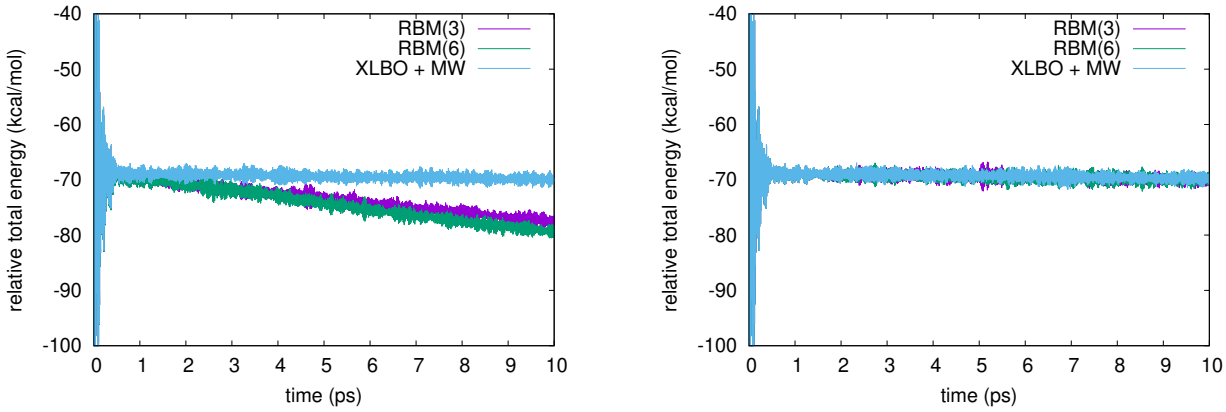

Figure 1: Total energy (kcal/mol) as a function of simulation time (fs) for 3HF comparing G-Ext(3), G-Ext(6) and XLBO with McWeeny purification, using a convergence threshold for the SCF algorithm of 10^{-5} (left panel) and 10^{-7} (right panel). The total energy was shifted of +505 000 kcal/mol for readability.

extrapolation, we can map the new density back to the original manifold, providing thus an idempotent density. The main element of novelty of the algorithm is that, by working on a tangent space, it allows one to use any linear extrapolation technique, while at the same time automatically ensuring the correct geometrical structure of the density matrix. As such, the technique presented in this paper can be seen as a simple case of a more general framework. Such a framework allows one to recast the problem of predicting a guess density by extrapolating information available from previous MD steps as a mapping between two vector spaces, i.e., the space of molecular descriptors and the tangent plane. This geometric approach can be seen as an alternative to extrapolating quantities derived from the density, such as the Fock or Kohn–Sham matrix, as proposed by Pulay and Fogarasi[4] and by Herbert and Head-Gordon.[5] However, the framework we developed, using molecular descriptors and a general linear extrapolation technique, can in principle be easily extended to such approaches.

That being said, our choices of both the molecular descriptor and of the extrapolation strategies are far from being unique. In recent years, molecular descriptors gained attraction within the rise of machine-learning (ML) techniques. Our choice, namely, using the Coulomb matrix, is only one of the many possibilities, and while being simple and effective, more ad-

vanced descriptors may be used and possibly improve the overall performances of the method. We also used a straightforward (stabilized) least-square interpolation of the descriptors at previous point to compute the extrapolation coefficients for the densities. This strategy is, again, simple yet effective. However, many other approaches can be used. In particular, ML techniques may not only provide a very accurate approximated map, but also benefit of a larger amount of information (i.e., use the densities computed at a large number of previous steps), further improving the accuracy of the guess. Improvements on the descriptors and extrapolation strategies are not the only possible extensions of the proposed method. A natural extension that is under active investigation is the application to the G-Ext guess to geometry optimization, for which the XLBO scheme cannot be used.

Overall, even the simple choices made in this contribution produced an algorithm that exhibits promising performances. In all our tests, the G-Ext method outperformed the well-established XLBO technique, especially for tighter SCF accuracies which may be relevant for post-SCF BOMD computations, including computations on excited-state PES. While we tested the method only for KS DFT, it can also be used for Hartree–Fock or semiempirical calculations. The main disadvantage of the proposed strategy with respect to the XLBO method is, however, the lack of time reversibility, which manifests itself as a lack of long-term energy conservation. In particular, for longer MD simulations, the total energy may exhibit a visible drift, which is something that the user must be aware of. In our test, the observed drift was relatively small and the use of a thermostat should be enough to avoid problems in practical cases, however, this is a clear, and expected, limitation of the proposed approach. We note that, using a tighter SCF convergence, which is also the case where the proposed method shows its best performances, produces an energy conserving trajectory, even starting from very noisy conditions. A time-reversible generalization of the G-Ext method is anyways particularly attractive, and is at the moment under active investigation.

Supporting Information Available

A JULIA template of the G-Ext algorithm is available at `https://github.com/epolack/GExt.jl`. The figure representing the total energy computation with an SCF convergence threshold of 10^{-6} for the molecule 3HF and formulas for the exponential and logarithm functions are available in the supplementary information.

Funding

Part of this work was supported by the French "Investissements d'Avenir" program, project ISITE-BFC (contract ANR-15-IDEX-0003).

References

(1) Fang, J.; Gao, X.; Song, H.; Wang, H. On the Existence of the Optimal Order for Wavefunction Extrapolation in Born-Oppenheimer Molecular Dynamics. *J. Chem. Phys.* **2016**, *144*, 244103.

(2) Alfè, D. Ab Initio Molecular Dynamics, a Simple Algorithm for Charge Extrapolation. *Comput. Phys. Commun.* **1999**, *118*, 31–33.

(3) Arias, T. A.; Payne, M. C.; Joannopoulos, J. D. Ab Initio Molecular-Dynamics Techniques Extended to Large-Length-Scale Systems. *Phys. Rev. B* **1992**, *45*, 1538–1549.

(4) Pulay, P.; Fogarasi, G. Fock Matrix Dynamics. *Chem. Phys. Lett.* **2004**, *386*, 272–278.

(5) Herbert, J. M.; Head-Gordon, M. Accelerated, Energy-Conserving Born–Oppenheimer Molecular Dynamics via Fock Matrix Extrapolation. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3269–3275.

(6) Hutter, J.; Parrinello, M.; Vogel, S. Exponential Transformation of Molecular Orbitals. *J. Chem. Phys.* **1994**, *101*, 3862–3865.

(7) VandeVondele, J.; Hutter, J. An Efficient Orbital Transformation Method for Electronic Structure Calculations. *J. Chem. Phys.* **2003**, *118*, 4365–4369.

(8) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

(9) Niklasson, A. M. N.; Tymczak, C. J.; Challacombe, M. Time-Reversible Born-Oppenheimer Molecular Dynamics. *Phys. Rev. Lett.* **2006**, *97*, 123001.

(10) Niklasson, A. M. N. Extended Born-Oppenheimer Molecular Dynamics. *Phys. Rev. Lett.* **2008**, *100*, 123004.

(11) Niklasson, A. M. N.; Steneteg, P.; Odell, A.; Bock, N.; Challacombe, M.; Tymczak, C. J.; Holmström, E.; Zheng, G.; Weber, V. Extended Lagrangian Born–Oppenheimer Molecular Dynamics with Dissipation. *J. Chem. Phys.* **2009**, *130*, 214109.

(12) Loco, D.; Lagardère, L.; Caprasecca, S.; Lipparini, F.; Mennucci, B.; Piquemal, J.-P. Hybrid QM/MM Molecular Dynamics with AMOEBA Polarizable Embedding. *J. Chem. Theory Comput.* **2017**, *13*, 4025–4033.

(13) Loco, D.; Lagardère, L.; Cisneros, G. A.; Scalmani, G.; Frisch, M.; Lipparini, F.; Mennucci, B.; Piquemal, J.-P. Towards large scale hybrid QM/MM dynamics of complex systems with advanced point dipole polarizable embeddings. *Chem. Sci.* **2019**, *10*, 7200–7211.

(14) Nottoli, M.; Mennucci, B.; Lipparini, F. Excited State Born-Oppenheimer Molecular

Dynamics through a coupling between Time Dependent DFT and AMOEBA. *Phys. Chem. Chem. Phys.* **2020**, *22*, 19532–19541.

(15) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433–14448.

(16) Niklasson, A. M. N. Next generation extended Lagrangian first principles molecular dynamics. *J. Chem. Phys.* **2017**, *147*, 054103.

(17) Niklasson, A. M. N. Density-Matrix Based Extended Lagrangian Born–Oppenheimer Molecular Dynamics. *J. Chem. Theory Comput.* **2020**, *16*, 3628–3640.

(18) Nottoli, M.; Bondanza, M.; Lipparini, F.; Mennucci, B. An enhanced sampling QM/AMOEBA approach: The case of the excited state intramolecular proton transfer in solvated 3-hydroxyflavone. *J. Chem. Phys.* **2021**, *154*, 184107.

(19) Polack, E.; Mikhalev, A.; Dusson, G.; Stamm, B.; Lipparini, F. An Approximation Strategy to Compute Accurate Initial Density Matrices for Repeated Self-Consistent Field Calculations at Different Geometries. *Mol. Phys.* **2020**, *118*, e1779834.

(20) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930–17955.

(21) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(22) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.

(23) Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

(24) Manzhos, S.; Dawes, R.; Carrington, T. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.* **2015**, *115*, 1012–1020.

(25) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, e1603015.

(26) Chmiela, S.; Sauceda, K.-R., Huziel E.and Müller; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.

(27) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(28) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(29) Goscinski, A.; Fraux, G.; Imbalzano, G.; Ceriotti, M. The role of feature space in atomistic learning. *Mach. learn.: sci. technol.* **2021**, *2*, 025028.

(30) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

(31) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B Condens. Matter* **2019**, *99*, 014104.

(32) Bachmayr, M.; Csanyi, G.; Drautz, R.; Dusson, G.; Etter, S.; van der Oord, C.;

Ortner, C. Atomic Cluster Expansion: Completeness, Efficiency and Stability. 2019; https://arxiv.org/abs/1911.03550.

(33) Edelman, A.; Arias, T. A.; Smith, S. T. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. Matrix Anal. Appl.* **1998-01-01**, *20*, 303–353.

(34) Zimmermann, R. Manifold Interpolation and Model Reduction. 2019; http://arxiv.org/abs/1902.06502.

(35) Bondanza, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. The Multiple Roles of the Protein in the Photoactivation of Orange Carotenoid Protein. *Chem* **2020**, *6*, 187–203.

(36) Bondanza, M.; Cupellini, L.; Faccioli, P.; Mennucci, B. Molecular Mechanisms of Activation in the Orange Carotenoid Protein Revealed by Molecular Dynamics. *J. Am. Chem. Soc.* **2020**, *142*, 21829–21841.

(37) Hashem, S.; Macaluso, V.; Nottoli, M.; Lipparini, F.; Cupellini, L.; Mennucci, B. From crystallographic data to the solution structure of photoreceptors: the case of the AppA BLUF domain. *Chem. Sci.* **2021**, accepted paper, doi: 10.1039/D1SC03000K.

(38) Becke, A. Density-Functional Thermochemistry. 3. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(39) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self — Consistent Molecular Orbital Methods . XII . Further Extensions of Gaussian — Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules Publishing Articles you may be interested in Self - consistent molecular or. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(40) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(41) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; andM. A. Robb, G. E. S.; Cheeseman, J. R.; Scalmani, G.; andG. A. Petersson, V. B.; Nakatsuji, H.; Li, X.; Marenich, A. V.; andJ. Bloino, M. C.; Janesko, B. G.; Zheng, J.; Gomperts, R.; andH. P. Hratchian, B. M.; Ortiz, J. V.; Izmaylov, A. F.; andD. Williams-Young, J. L. S.; Ding, F.; Lipparini, F.; Egidi, F.; andB. Peng, J. G.; Petrone, A.; Henderson, T.; Ranasinghe, D.; andJ. Gao, V. G. Z.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; andR. Fukuda, K. T.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; andH. Nakai, O. K.; Vreven, T.; Throssell, K.; J. A. Montgomery, J. a. E. P.; Ogliaro, F.; Bearpark, M. J.; andE. N. Brothers, J. J. H.; Kudin, K. N.; Staroverov, V. N.; andR. Kobayashi, T. A. K.; Normand, J.; Raghavachari, K.; andJ. C. Burant, A. P. R.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; andM. Klene, J. M. M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; andK. Morokuma, R. L. M.; Farkas, O.; Foresman, J. B.; ; Fox, D. J. Gaussian Development Version, Revision J.16. 2020; Gaussian, Inc., Wallingford CT, 2020.

(43) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289.

(44) Interfacing to Gaussian 16 (v2). `https://gaussian.com/interfacing/`, Last accessed: 1st June 2018.

(45) McWeeny, R. Some Recent Advances in Density Matrix Theory. *Rev. Mod. Phys.* **1960**, *32*, 335–369.

Grassmann extrapolation of density matrices for Born-Oppenheimer molecular dynamics

Supplementary information

Étienne Polack Geneviève Dusson Benjamin Stamm
Filippo Lipparini

September 17, 2021

Supplementary figure

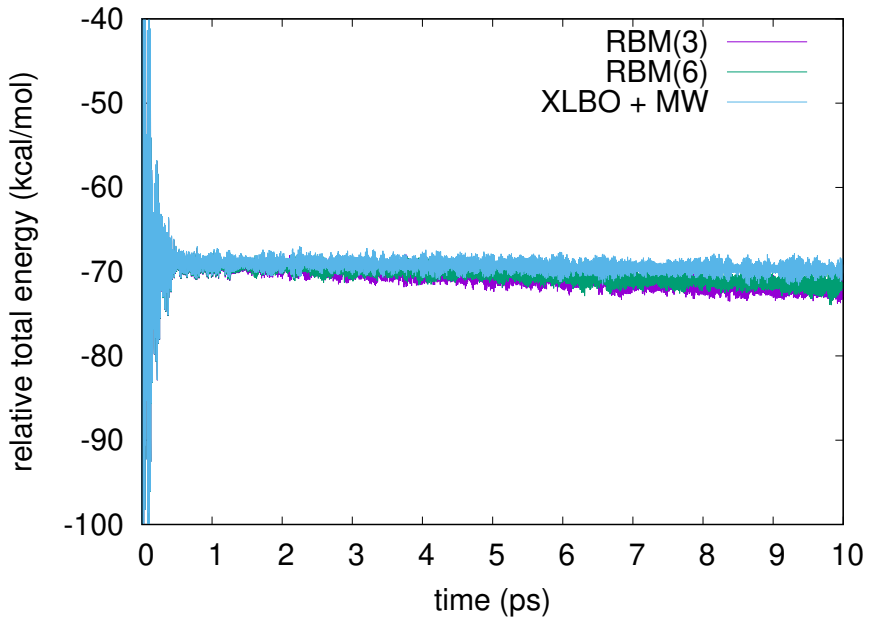

Figure 1: Total energy (kcal/mol) as a function of simulation time (fs) for 3HF comparing G-Ext(3), G-Ext(6) and XLBO with McWeeny purification, using a convergence threshold for the SCF algorithm of 10^{-6}. The total energy was shifted of $+505\,000$ kcal/mol for readability.

Grassmann Exponential and Logarithm maps

The Grassmann manifold is a differential manifold and, for any given $D_0 = C_0 C_0^\mathsf{T} \in \mathcal{G}r(N,\mathcal{N})$ with $D_0 := D_{\mathsf{R}_0}$ and $C_0 := C_{\mathsf{R}_0}$ for fixed R_0, the tangent space is characterized by

$$\mathcal{T}_{D_0} = \left\{ \Gamma \in \mathbb{R}^{\mathcal{N}\times N} \,\middle|\, C_0^\mathsf{T}\Gamma = 0 \right\} \subset \mathbb{R}^{\mathcal{N}\times N}. \tag{1}$$

Note that the tangent space is a linear space. One can then introduce the Grassmann exponential which maps tangent vectors on $\mathcal{T}_{D_0}$ to the manifold $\mathcal{G}r(N,\mathcal{N})$ in a locally bijective manner around D_0. Indeed, it is not only an abstract tool from differential geometry, but it can be computed in practice involving the matrix exponential. By complementing C_0 with orthonormal columns to obtain $(C_0, C_\perp) \in O(\mathcal{N})$, where $O(\mathcal{N})$ denotes the group of orthogonal matrices of dimension $\mathcal{N}\times\mathcal{N}$, and $\Gamma \in \mathcal{T}_{D_0}$ we have

$$\mathrm{Exp}_{D_0}(\Gamma) = CC^\mathsf{T}, \qquad C = \left(C_0, C_\perp\right) \exp\begin{pmatrix} 0 & -B^\mathsf{T} \\ B & 0 \end{pmatrix} \mathsf{I}_{\mathcal{N},N}. \tag{2}$$

Here, exp denotes the matrix exponential function, the matrix $B \in \mathbb{R}^{(\mathcal{N}-N)\times N}$ contains expansion coefficients of columns of Γ in a span of columns of $C_\perp$ such that $\Gamma = C_\perp B$ and $\mathsf{I}_{\mathcal{N},N} = (\mathsf{I}_N, 0)^\mathsf{T} \in \mathbb{R}^{\mathcal{N}\times N}$ are the first N columns of the $\mathcal{N}\times\mathcal{N}$ identity matrix. As described in [1, 2], the Grassmann exponential can then be equivalently expressed by

$$\mathrm{Exp}_{D_0}(\Gamma) = CC^\mathsf{T}, \qquad C = \left[C_0 V_e \cos(\Sigma_e) + U_e \sin(\Sigma_e)\right] V_e^\mathsf{T}, \tag{3}$$

by means of a singular value decomposition (SVD) of the matrix $\Gamma = U_e \Sigma_e V_e^\mathsf{T}$.

The inverse function is the so-called Grassmann logarithm Log_{D_0} (see, e.g., [1, 2]) which maps any $D = CC^\mathsf{T} \in \mathcal{G}r(N,\mathcal{N})$ in a neighborhood of D_0 to the tangent space $\mathcal{T}_{D_0}$ by

$$\mathrm{Log}_{D_0}(D) = U_\ell \arctan(\Sigma_\ell) V_\ell^\mathsf{T}, \tag{4}$$

using the following SVD decomposition

$$U_\ell \Sigma_\ell V_\ell^\mathsf{T} = L \qquad \text{with} \qquad L = C\left(C_0^\mathsf{T} C\right)^{-1} - C_0. \tag{5}$$

References

[1] Alan. Edelman, Tomás A. Arias, and Steven T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998-01-01.

[2] Ralf Zimmermann. Manifold interpolation and model reduction, 2019. http://arxiv.org/abs/1902.06502.