Systematically Improvable and Locality Accelerated Enzymatic Reactivity Modeling: Toward Chemical Accuracy at Affordable Cost

Dénes Berta,[†,‡,¶] József Csóka,[†,‡,¶] Gyula Samu,[†,‡,¶] and Péter R. Nagy[*,†,‡,¶]

†*Department of Physical Chemistry and Materials Science, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

‡*HUN-REN–BME Quantum Chemistry Research Group, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

¶*MTA–BME Lendület Quantum Chemistry Research Group, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

E-mail: nagy.peter@vbk.bme.hu

Abstract

Quantum mechanics/molecular mechanics (QM/MM) is the cornerstone of computational enzymology. Herein, we address an outstanding challenge in QM/MM, namely, simultaneous access to accurate QM methodology and a large QM subsystem at an affordable computational cost. First, local natural orbital (LNO) based CCSD(T) is employed for chemically accurate energetics and references for choosing density functional theory (DFT) models. Next, reliable hybrid DFT methods are selected, with large QM subsystem selections, suitable also for reaction barriers. Then, quantum embedding, especially accelerated via our recent local embedded subsystem (LESS) approach, is used to reduce the cost of DFT calculations to a few core hours even with large QM

sizes up to ca. 400 QM atoms. By combining these advanced methods, we propose a Locality Accelerated (by LESS and LNO) and Systematically Improvable (LASI) scheme for QM/MM simulations. It benefits from the strengths of converged QM size in its DFT component, affordability for many configurations via quantum embedding, and, if needed, CCSD(T) accuracy for energetics. The protocol is validated through the study of challenging, representative, and clinically relevant enzyme-catalyzed phosphate hydrolysis. Based on these results, we establish generally applicable guidelines to set up the components of the LASI protocol. The flexibility and affordability of LASI, both in large scale QM or QM/MM context, make it broadly applicable for the predictive computational description of enzyme reactivity and beyond.

1 Introduction

The computational description of enzyme catalysis provides a powerful framework for elucidating complex biochemical reactions. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) has emerged as the *de facto* methodology for simulating reactions within biomolecular environments, becoming essential in mechanistic studies and in understanding interaction driving drug design.[1–3] The comparison of experimental and computational properties hinges on the calculation of (relative) free energies, as these thermodynamic quantities are the basis for robust validation.[4] A validated simulation protocol transitions from a descriptive tool to a predictive one, capable of guiding the rational design of new molecules or materials.

Obtaining accurate kinetic and thermodynamic properties necessitates a careful compromise between multiple coupled aspects. Here, we focus on three of central importance: (A1) the accuracy of the QM method, (A2) the size of the subsystem treated by that QM method, and (A3) the exhaustiveness of the statistical sampling technique.[5] Our objective is to advance QM/MM methodologies, well-converged in QM accuracy (A1) and QM size (A2), in a way that when combined with enhanced sampling applications (A3), free energy simulations will still be affordable. Although the choice of each simulation parameter introduces inherent and coupled uncertainties, recent advances enable us to approach them via systematically improvable methods. The convergence of properties with regard to a single aspect (either QM accuracy, QM size or sampling) have been established in various cases, however combining the individually converged levels remains prohibitively expensive.

Regarding A1, gold-standard, wavefunction-based methods enable chemical accuracy (ca. 1 kcal/mol). These are becoming increasingly affordable by themselves and can also provide references for selecting reliable, e.g. hybrid level Density Functional Theory (DFT) models. Concurrently, for A2, many kinds of computed properties exhibit a slow convergence with respect to the QM region size. In turn, calculations for such converged QM sizes have been limited to a small number of structures with hybrid DFT. On top of these challenges, the

rigorous demands of statistical thermodynamics necessitates even more compromise in A1–A2. Specifically, the high computational cost of extensive sampling, which is required for free energy convergence, often forces practical compromises both on the DFT model, basis set and the QM region size, thereby introducing limitations in accuracy.

Coupled cluster (CC) with single, double and perturbative triple excitation [CCSD(T)] methods[6] represent the gold-standard regarding the accuracy (A1) of quantum methods, while the best and well-selected DFT functionals may also approach chemical accuracy for a specific reaction. Local correlation-based CCSD(T) calculations, such as our local natural orbital (LNO), are now accessible relatively routinely for systems comprising several hundred atoms, providing excellent energetic predictions.[7] However, CCSD(T) forces are not available yet for large molecules, limiting their use in free energy sampling. Using such high-level references, accurate DFT functionals can be rationally selected for potential energy optimizations, enabling identification of key transition and intermediate states.[8]

The convergence of computed properties with respect to the size of the QM region is a critical consideration (A2).[9,10] Enlarging the QM region results in more complete incorporation of quantum interactions, like electronic polarization, charge transfer, long-range dispersion and their coupling, which enhances accuracy but increases computational demand steeply. Practical limitations arise from the computational cost, thus requiring careful selection of the QM region that balances capturing essential chemistry with tractability. Studies of the convergence of calculated properties suggest that the QM regions may often need to contain several hundreds of atoms for properties to converge.[11–15] However, in practice, selections are often compromise due to the computational cost of the chosen QM methodology, and stop short of size-convergence (A2) at around 100 QM atoms.[5]

Free energy methods to characterize reaction thermodynamics in QM/MM contexts typically involve sampling techniques (A3) to compute measurable properties through statistical means.[16] For reactions, this is further complicated by regions in the free energy landscape that are visited with a low probability (e.g. towards transition states). Simulations thus rely

on enhanced sampling methods, like umbrella sampling, metadynamics, thermodynamic integration or the string method.[17,18] Despite their successes, these approaches face limitations because convergence requires exhaustive sampling of the high-dimensional phase space, and the computational cost of ab initio QM/MM calculations restricts the accessible simulation timescales and system sizes.[19,20] Consequently, their combination with accurate QM levels of theory and sufficiently large QM regions are hampered.[21] The common choices are for converged statistics (A3), most often rely on classical, semiempirical, or recently also machine learned force field methods.[22–26]

We demonstrate these limitations and the capabilities of our proposed approach to overcome them on the example of the GTP hydrolysis catalyzed by the extensively studied small GTPase Ras. The clinical relevance of Ras is in its high rate of mutation in cancer cell lines, also associated with poor prognosis.[27,28] GTPases cleave the terminal phosphate group of the anhydride chain in GTP (Figure 1). Breaking or forming of P-O bonds is a common in biochemical reactions, as phosphates are an essential building block of living organisms.[29,30] The mechanism of Ras GTPase has been extensively studied.[31] Building on this, we use five stationary points (Figure 1) to represent the general base catalyzed mechanism published in Ref. 20. The heavily polarized active site is challenging for frequently used DFT

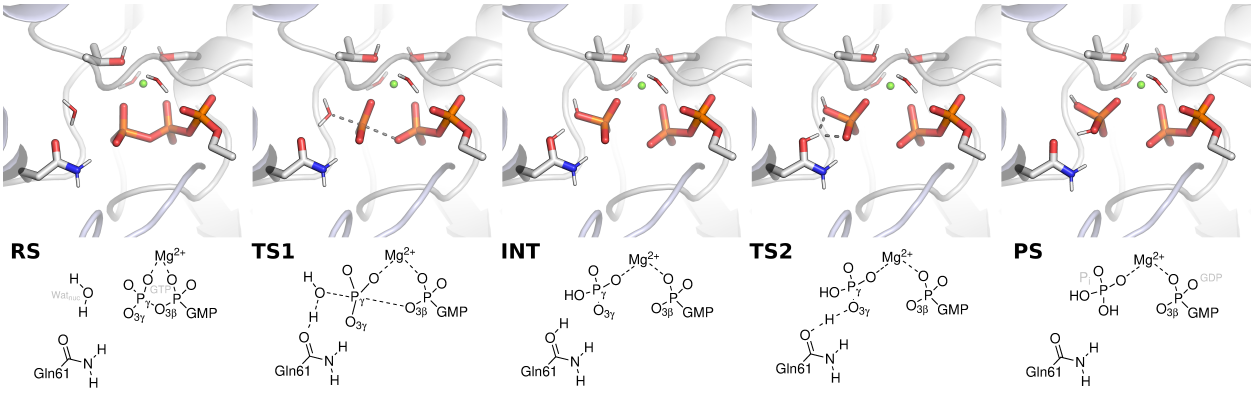

Figure 1: Stationary points along the Gln61 assisted GTP hydrolysis mechanism of the Ras GTPase: RS: reactant state, TS1: transition state 1, INT: intermediate, TS2: transition state 2, PS: product state. Forming and breaking bonds are depicted in gray dashes. The guanosine of the GTP and aliphatic hydrogens are omitted for clarity.

methods,[32–34] especially because it includes numerous explicitly charged residues, such as GTP^{4-}/GDP^{3-} coordinated Mg^{2+} ion. The S_N2 character of the phosphate cleavage poses further difficulty for practical density functional approximations. The associated S_N2 TS is known to require at least hybrid DFT, driving the cost of the QM/MM calculations. The phosphate cleavage is followed by a proton transfer step, making this a representative real-life system to study the accuracy and limitations of QM/MM modeling describing phosphate biochemistry.[31]

Herein we propose improvements to QM/MM approaches by extending the use of cutting-edge computational techniques to model biochemical reactivity. We specifically focus on the coupling and systematic convergence of QM methods and QM region selection, toward reaching chemical accuracy in A1 and A2. Moreover, we do so by also working with the computational budget restraints required for enhanced sampling applications (A3).

To that end, the following advances and demonstrations are presented: The convergence of reaction barriers and energies are studied against QM region sizes up to several hundred atoms. From this we identify general selection principles that help achieving converged results with less resources. We demonstrate the applicability of well-converged LNO-CCSD(T) calculations for QM regions up to about 400 atoms with accessible resources. Such reference values are used here to select robust and accurate DFT methodologies including a sufficiently large basis set for the studied phosphate chemistry. Crucially, we study the connection between convergence with the QM method and the region size to select the appropriate model parameters. We compare multiple ways of accelerating DFT calculations, of which our recently developed local embedded subsystem (LESS) framework stands out.[35] In LESS, we start with a quantum mechanical embedding approach, which in the present context entails hybrid DFT for the subsystem involved in chemical change and GGA DFT for an extended protein environment that still cannot be accurately modeled via MM. LESS exploits that the high level hybrid DFT computation can be greatly accelerated by a strategic selection of not only the high level atoms and occupied orbitals (like in previous quantum embedding vari-

ations) but also the atomic and density fitting basis functions.[35] We demonstrated recently that LESS practically achieves hybrid DFT accuracy at about twice the cost of GGA DFT.[35] Our approach bears two recurring characteristics: first, it relies on the performance acceleration via exploiting locality of the chemical process. LESS achieves acceleration for accurate DFT calculations, while LNO-based CC provides affordable chemical accuracy. Second, it builds on systematic improvability in terms of the employed local and embedding approximations as well as the converging series of basis set and QM subsystem size choices. Thus besides useful defaults, we present a process to select accurate and efficient methodologies tailored for different simulation goals.

2 Methods

2.1 Locality Accelerated, Systematically Improvable (LASI) Enzymatic Thermochemistry

We propose a relative free energy scheme, LASI for short, that is conscious of both systematic improvability and affordability of different methodologies, aiming at the best possible accuracy and transferability allowed by state-of-the-art methods. For a given system, we denote the most accurate model that is computationally compatible (e.g. can compute gradients) and fast enough to be used jointly with sampling (S) techniques as $\mathrm{DFT_S}$. In other words, $\mathrm{DFT_S}$ embodies optimum for statistical measures and consequently calculation of entropic contributions. Hence free energy simulations comprising of tens of thousands of points have to be possible on the level of $\mathrm{DFT_S}$, from which entropic contributions are calculated. If required, moving towards the models sizes and/or QM levels that are closer to the converged results, we propose to find the most quantitatively accurate method $\mathrm{DFT_E}$. It can be applied to identify reaction paths still requiring the evaluation of atomic forces in hundreds of points, but this $\mathrm{DFT_E}$ model mainly operates on the PES due to its higher cost. $\mathrm{DFT_E}$ puts the focus on A1 and A2 optimization as it is intended for the evaluation of a smaller

number of configurations. For a more limited number of structures, such as the comparison of stationary points or a discretely sampled reaction path, converged LNO-CCSD(T) is affordable. Using this as a correction to $\mathrm{DFT_E}$ energies ($\Delta\Delta E_{\mathrm{CC}}$) improves the energetics toward chemical accuracy.

Therefore the proposed free energy scheme (with all terms calculated in a QM/MM manner as electrostatically embedded into an MM environment) is as follows:

$$\Delta G = \Delta E_{\mathrm{DFT_E}} + [\Delta\Delta E_{\mathrm{CC}}] - T\Delta S_{\mathrm{DFT_S}}\,. \tag{1}$$

If it is affordable, it is simpler to use the same $\mathrm{DFT_S}$ and $\mathrm{DFT_E}$ model. We keep them separated, as $\mathrm{DFT_E}$ might be still expensive for practical free energy simulations at the level of sufficiently converged DFT level, QM regions and basis sets. Next, we overview the theoretical and technical developments that we leverage in determining the specific settings for LASI components $\mathrm{DFT_E}$, $\mathrm{DFT_S}$, and $\Delta\Delta E_{\mathrm{CC}}$, while keeping their respective cost requirements in mind.

2.2 Local Coupled-Cluster Calculations

For gold-standard references, well-converged CCSD(T) electronic energies are computed, using the efficient local natural orbital (LNO)[7,36–39] method of the MRCC program package.[40–42] Extrapolation toward the complete basis set (CBS) and the local approximation free (LAF)[38] limits of LNO-CCSD(T) are both performed. Therefore the reference composite energy scheme is written as follows:

$$E^{\mathrm{CBS}(X,X+1)}_{\text{N-T LNO-CCSD(T)}} = E^{\mathrm{CBS}(X,X+1)}_{\mathrm{Normal}} + E^{X}_{\text{N-T}} - E^{X}_{\mathrm{Normal}} \tag{2}$$

where X stands for the cardinal number in the used atomic orbital (AO) basis sets, N and T refer to the *Normal* and *Tight* local correlation thresholds and N-T denotes the extrapolation from these toward the LAF limit. This protocol in Eq. 2 was found accurate and affordable

across a broad range of chemical applications and its validity can always be checked system-specifically, like we do below.[7] The basis set convergence is further accelerated by adding a density-based basis set correction (DBBSC) to the LNO-CCSD(T) values.[39] DBBSC provides a short-range DFT-based correction for the incompleteness of the specific finite AO basis set used for wavefunction methods. This is achieved via a finite basis electron-electron interaction formulation that uses the density of the system.[43,44] Thus DBBSC corrects for the short-range finite basis error that is dominantly responsible for the slow basis set convergence of wavefunctions methods. Our LNO-based DBBSC implementation provides asymptotically linear-scaling and very low overhead for the rate determining step of DBBSC, i.e. the range-separation function of the DFT correction that adapts to the spatial inhomogeneity of the basis-set incompleteness.[39]

We establish the convergence of relative energies regarding both basis sets and thresholds on smaller QM regions and carry out LNO-CCSD(T)/CBS calculations up to the QM size used in DFT_E calculations (see Figure 4 and section S3 for more details). The accuracy of such references are excellent for benchmarking purposes, and their effectiveness enables their usage in Eq. 1 for 10s–100s of configurations along the reaction coordinate. Moreover, the approach fits into our overall strategy, as the basis set, LNO threshold, and the QM size can all be systematically converged to chemical accuracy. An additional benefit is that robust error estimates can be assigned to monitor the level of convergence.[7,38]

2.3 Speeding up DFT Calculations via Huzinaga Quantum Embedding

When modeling reactions, lower rung DFT functionals, like GGAs or mGGAs are known to overstabilize some transition states, e.g. due to self-interaction error (SIE), like for **TS1** in Figure 1.[8] The chosen enzymatic reaction, representing features typical in biochemistry, also involves multiple bond breaking/forming events. The environment is highly polarized by numerous explicitly charged moieties, posing further challenges for DFT models. Therefore

hybrid or higher level DFT methods are required, setting a computational bottleneck in QM/MM, especially limiting free energy simulations.

To speed up the DFT calculation, we advance the utility of quantum embedding methods for enzyme modeling.[45–47] Here, in addition to QM/MM, we split the QM calculation into two QM layers. This enables fast evaluation by limiting the expensive but accurate high-level (HL) QM description to the chemically active subsystem (A). Among the numerous QM-QM embedding schemes, our implementation is an improved variant of projection-based embedding formalisms,[48] using the Huzinaga-equation.[35,49] The total energy of the QM-QM system is written as

$$E = E_{\mathrm{LL}}[\boldsymbol{D}] - E_{\mathrm{LL}}[\boldsymbol{D}^A] + E_{\mathrm{HL}}[\tilde{\boldsymbol{D}}^A] + \mathrm{Tr}\left\{(\tilde{\boldsymbol{D}}^A - \boldsymbol{D}^A)\frac{\partial E_{\text{LL-HL}}}{\partial \boldsymbol{D}^A}\right\} , \tag{3}$$

where LL and HL stand for the low- and high-level models; $\boldsymbol{D}$ and $\boldsymbol{D}^A$ refer to the LL density of the full and the active subsystem, respectively, while $\tilde{\boldsymbol{D}}^A$ denotes the HL density calculated for the active subsystem. The last term of Eq. 3 is a first order correction in the Huzinaga-embedding energy using the active subsystem density difference and

$$E_{\text{LL-HL}} = E_{\mathrm{LL}}[\boldsymbol{D}] - E_{\mathrm{LL}}[\boldsymbol{D}^A] + E_{\mathrm{HL}}[\tilde{\boldsymbol{D}}^A] . \tag{4}$$

The separation of the HL and LL subsystems creates an opportunity to choose a more efficient DFT method and smaller basis set for the LL environment, while retaining the HL description for local processes in the active layer. In practice, first, the LL calculation is performed for the entire QM system, then the occupied molecular orbitals (MOs) are localized. The MOs residing on the subsystem atoms are selected as active MOs (orange halos in Figure 2), based on Mulliken population analysis[48,49] or via other methods.[50,51] From the practical perspective, the user only needs to select the active atom list, the MO selection is automated. Then, the high-level MOs are optimized in an embedding potential with the HL SCF method, while the environment MOs are kept frozen.

To further accelerate the HL computation, we have recently introduced the local embedded subsystem (LESS) framework, which decreases the size of the AO basis set, as well as the auxiliary basis set when density fitting (DF) is used (yellow halos in Figure 2). This yields asymptotically constant cost for the HL part at the limit of much larger environment around a finite active system. Consequently, only those AO and DF basis functions are used for the HL calculation of the embedded subsystem that are important to accurately expand the active MOs and Huzinaga embedding quantities needed for the HL part. Therefore the LESS approximations can drastically reduce the number of DF integrals, which in turn can be stored in memory (in-core) for the HL calculation. The calculation of analytical gradients are already available in our Huzinaga-type embedding implementation.[52] We are actively working on extending that to the LESS scheme to exploit the associated speed-up in optimizations and MD simulations.

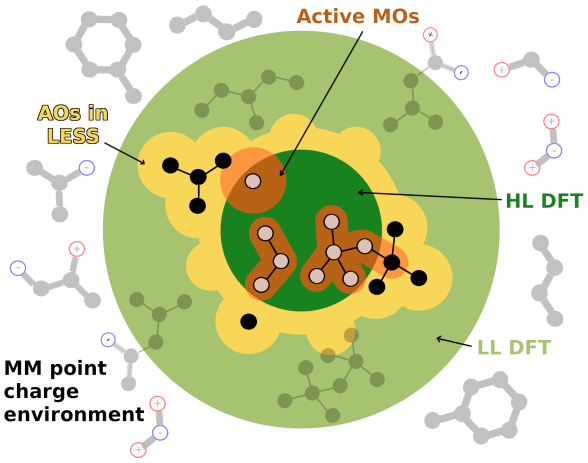

Figure 2: Schematic representation of the LESS framework, inside an electrostatically embedded MM environment (gray balls and sticks with charges). The high-level calculation of the active subsystem (dark green) comprised of the active atoms (hallow balls) only utilizes the MOs (orange spheres) and the AO and auxiliary functions (yellow halo). The low-level calculation of the full QM system is depicted in light green.

We previously demonstrated that LESS is universally applicable for various catalytic (homogeneous, surface and enzymatic) reactions to a similar effect: accuracy of a QM_{HL} method obtained at the cost of 2–3 QM_{LL} calculations due to the 50–100-fold acceleration of

the QM_{HL} part.[35] Based on those diverse cases, reasonable defaults are also available for basis truncation algorithms. Importantly, one can always systematically converge to the QM_{HL} results by improving the QM_{HL} region definition, as well as the AO or DF basis selecting thresholds.

An alternative for speeding up hybrid DFT calculations is our local DF (LDF) algorithm.[53–55] Here a set of atoms (a domain) is automatically selected for each localized MO, done so that the auxiliary functions centered on the selected atoms are used in the exchange term for DF. The LDF approach can significantly speed up the hybrid DFT calculation without relying on embedding. The main difference is that the HF exchange contribution is evaluated for all MOs in LDF, while with LESS Huzinaga embedding, the HF exchange evaluation is restricted to the reactive regions.

2.4 Selection of the QM Region in QM/MM

In an enzymatic environment, the definition of QM region of QM/MM will inevitably result in cutting covalent bonds, invoking treatment of free valences by link atoms or localized single electron orbitals.[56,57] The general advice is to make cuts at apolar single bonds, typically carbon-carbon, whenever possible.[58] The immediate chemical environment of a reaction is generally straightforward from the reaction in question and mechanistic ideas. In the context of a reaction, especially with multiple elementary steps, the set of important residues in a product state are not necessarily identical to that of the reactant state. Consequently, with a preliminary reaction path, selections can be made for multiple structures (e.g. reactant, product states), then the union of those could be used in downstream simulations.[59] Including the reactive moieties and their close contacts often yield a QM region around 100 atoms.[60] Due to the associated cost of at least hybrid DFT, many studies do not exceed this consideration, while key energetic properties may not be converged to chemical accuracy.

While there are algorithms proposed to aid the definition of QM regions, e.g. based on electrostatics,[61,62] forces[63] or contact maps,[64] the QM region selection of an arbitrary active

site does not have a universally accepted technique.[57] Selection methods are also developed for building QM cluster models, with the potential to be extended for QM/MM simulations.[65,66] While in most studies with less (50–100) QM atoms, the optimization of the QM atom selection is somewhat simpler, this problems becomes more pressing as one aims at convergence with the QM size and therefore includes hundreds of QM atoms. Namely, the complexity of optimal QM atom list selection rapidly grows with the spatial extension of the QM region and at the same time, the chemical-knowledge-based selection becomes ambiguous. Therefore the considerations that current (semi-)automated selection algorithms use for building a QM/MM interface is and will be even more valuable as larger QM selections become more affordable.[67] Consequently, one may anticipate further focus and progress on automated selection algorithms in the future, especially applied in high-throughput screening procedures. Notably, if the constructed model is intended for extended reaction path optimization and free energy simulation, thorough manual optimization of the QM region is justified by the downstream gain in computational cost. In other applications, hand-optimized QM selections may require more time investment than the benefit from QM/MM time reduction. In such cases, one could more easily afford simplified or automated QM region selector algorithms because of the acceleration offered by the LESS approach for larger total QM selections.

Here, we use a topologically extended distance-based QM selection algorithm as a starting point of our QM region optimization. The extension starts from a manually selected core atom list, which can be as small as a single atom, making sure that we have appropriate focus on where the reaction happens. Atoms within a defined radius are then added, unlike in some selection methods, which decide about the inclusion of residues as a whole. The thus created atom list is then iteratively extended along the covalent bonds until it encounters a cutable bond. The list of bonds we allow to be cut at the interface is defined for the conventional protein atom types and can be extended in our implementation. The selection algorithm also removes topologically close link atoms: if the MM-side atom is the same at

two cutting events, the QM region is automatically extended, thereby decreasing number of cut bonds at the QM/MM interface. Another important feature is the option to include multiple structures for the selection, which is ideal for considering geometry changes along a preliminary reaction path. The outlined QM atom selection script and its documentation is openly available on GitHub.[68]

3 Results

In this section, we present the convergence of reaction energies and barriers both with the increasing size of the QM region and systematically improved computational methods against gold-standard LNO-CCSD(T) references. We also study the connection between these convergences, determining model details for different cost levels of the LASI scheme discussed in Eq. 1. Finally, we analyze the effect of multi-layer QM methods on the accuracy of the calculated relative energies. Computational details are collected in section 6.

3.1 Initial Optimization of the QM Atom List

The minimal chemically intuitive QM region from the previous QM/MM studies on the selected reaction was 101 atoms, including link atoms.[20] Further trimming of this resulted unrealistic increase in the barrier height of the second step (Figure 3C, right, green 'chemical'). We first considered radial extensions in the range of 3 to 9 Å radius from four reactive atoms: the γ-phosphorus, the oxygen of the nucleophilic water, the oxygen atom of the leaving group and the Mg^{2+}(red halos in Figure 3A). Relative energies of the stationary points converge slowly with radial selections, convergence within 1 kcal/mol is only reached above 431 QM atoms (blue series in Figure 3C and S1). The convergence of the overall relative QM/MM energies above 400 atoms can also be observed in the slow convergence of the dispersion correction or the disappearance of relative MM contribution (Figure S2). Interestingly, the two reaction steps exhibit very similar overall barrier heights (Figure 3C inset), which makes

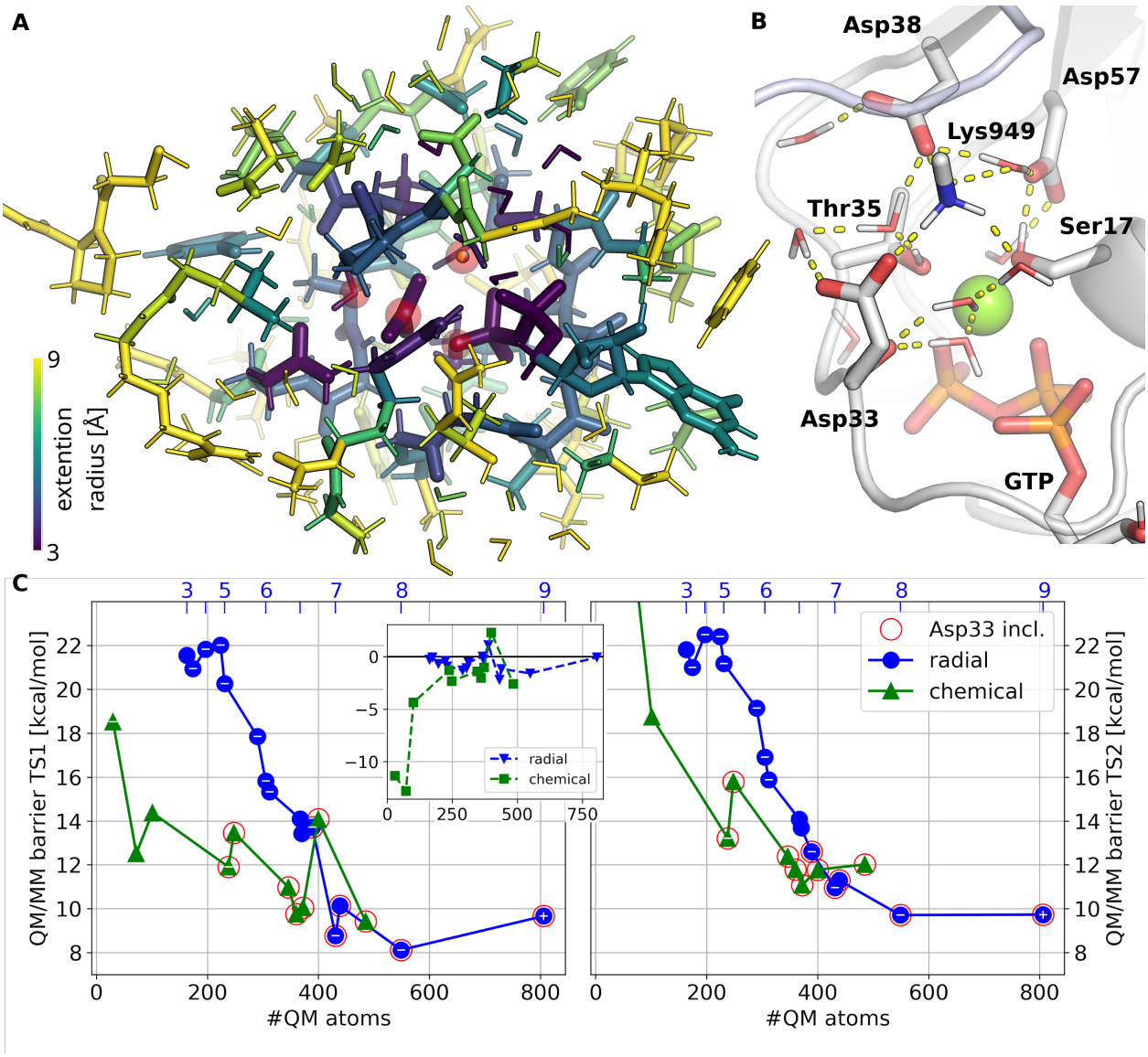

Figure 3: **A** Series of QM selections (30–806 QM atoms, depicted in purple-yellow color scale) extended from four central atoms (Mg^{2+}, γ-phosphorus, the oxygen of the nucleophilic water and the $\beta - \gamma$ bringing oxygen) highlighted in red. **B** Hydrogen-bonding network near the Mg^{2+} binding site. **C** Convergence of the QM/MM energies (B3LYP-D3/6-31+G*/CHARMM36m) along two series of QM selections: blue circles: 'radial' extension from the reaction center, green triangles: selections prioritizing explicit charges, denoted as 'chemical'. Selection radii are depicted for the radial series in the secondary x axis. Formal charge is indicated on the markers if not neutral. Convergence is represented by the two barriers (TS1–RS, left and TS2–RS, right) and their relative energies (TS1–TS2, inset). Selections including Asp33 (and the complete network shown in **B**), are highlighted by red circles. Other relative energies are presented in Figure S1.

determining the rate-determining step difficult. This highlights the need for well-converged results especially when comparing transition states of different nature, otherwise one may arrive at qualitatively wrong conclusions regarding the mechanistic details.

To optimize the QM region, we examined the individual exclusion of several residues (or part of thereof) at the edge of QM regions. The moieties should be kept in the QM region if the relative energies change significantly upon their exclusion (Figure S4). The removal of explicitly charged residues had the largest effect on the relative energies. Especially important is Asp33, which is part of a dense hydrogen-bonding network around the Mg^{2+} binding site (Figure 3B).[69] This suggests that the classical, non-polarized representation from the protein force field does not adequately describe the interactions between and around explicit charges.[70] Based on the importance of polar contacts observed in the convergence study, one may reasonably suspect that polarizable embedding in QM/MM would accelerate the convergence shown in Figure 3. However, the associated cost of the extra self consistent optimization of multipoles in such setup were found to be significant (7-11 times vs electrostatic embedding).[71] Considering the favorable scaling of LESS (see section 4.1), the inclusion of residues with significant polarizability near the active site to the LL environment is an affordable and universal solution not requiring parametrization. Interestingly, the removal of water molecules at the edge of this hydrogen bonding network, but far from the reaction site appears to have little effect (Figure S4). Although this extensive QM region optimization is not necessary for establishing convergence, it is justified by the cumulative gains from downstream calculations, that is one of the aims of the LASI approach.

Crucially, these selections allowed the fragmentation of residues at apolar C-C bonds. Let us note that the combination of radial extension with the selection of complete residues only, often used in automated selection algorithms, in average would constitute a 45% increase in the size of the QM region (Figure S5). In turn this larger QM size would lead to ca. five-fold increase in the computational cost at hybrid DFT level. We also studied whether the number of covalent cuts made at the interface (and the addition of the corresponding link atoms)

has any implication for the convergence. We found that neither the number nor the ratio of link atoms (vs the full QM size) alone show significant correlation with the size-related error. Together with the illustrated importance of the polar networks, this suggests that the quality of the (covalent) cuts at the interface and the insufficiency of the MM model for key polar groups are more important than the size of the QM/MM interface.[72]

Based on these observations, we created selections with minimal extension radii (2–3 Å) around the explicitly charged and key reactive atoms (see section S1 in the SI). The new series exhibits faster convergence to the asymptotically large QM regions (green series in Figure 3C). We select two QM regions, with 238 and 372 atoms, that are considerably closer to convergence than the general trends. These are the selections for building the DFT_S and DFT_E models, respectively, because good convergence in the difference of barriers is reached at 238 atoms, while the same for all relative energies is achieved with 372 atoms. Let us highlight that these selections are optimized for accuracy, and would probably be beyond the affordable cost without further acceleration in section 3.3.

3.2 Selecting a QM Method against LNO-CCSD(T) References

To establish the QM references, the convergence with local approximations and basis sets were assessed in LNO-CCSD(T)/CBS calculations for three representative QM regions (101, 238 and 372 QM atoms). The basis set dependence exhibit typical convergence patterns (Figures 4 and S6–S7), which is accelerated by the application of the DBBSC. Basis set corrected values are close to chemical accuracy (1 kcal/mol) already with triple-ζ basis sets (triangles in Figure 4), both with and without diffuse functions. The CBS(T,Q) references are also in agreement with both the cc-pVXZ and aug-cc-pVXZ series for the 101 (Figure 4A) and 238 (Figure 4B) QM atom cases. Consequently, we omitted the most expensive calculations for the 372-atom model with augmented basis sets. Considering the feasibility and accuracy of the CBS(T,Q) calculation with DBBSC, we recommend it as a widely applicable reference method for similar (bio)chemical applications. Namely, LNO-CCSD(T)+DBBSC for the 101

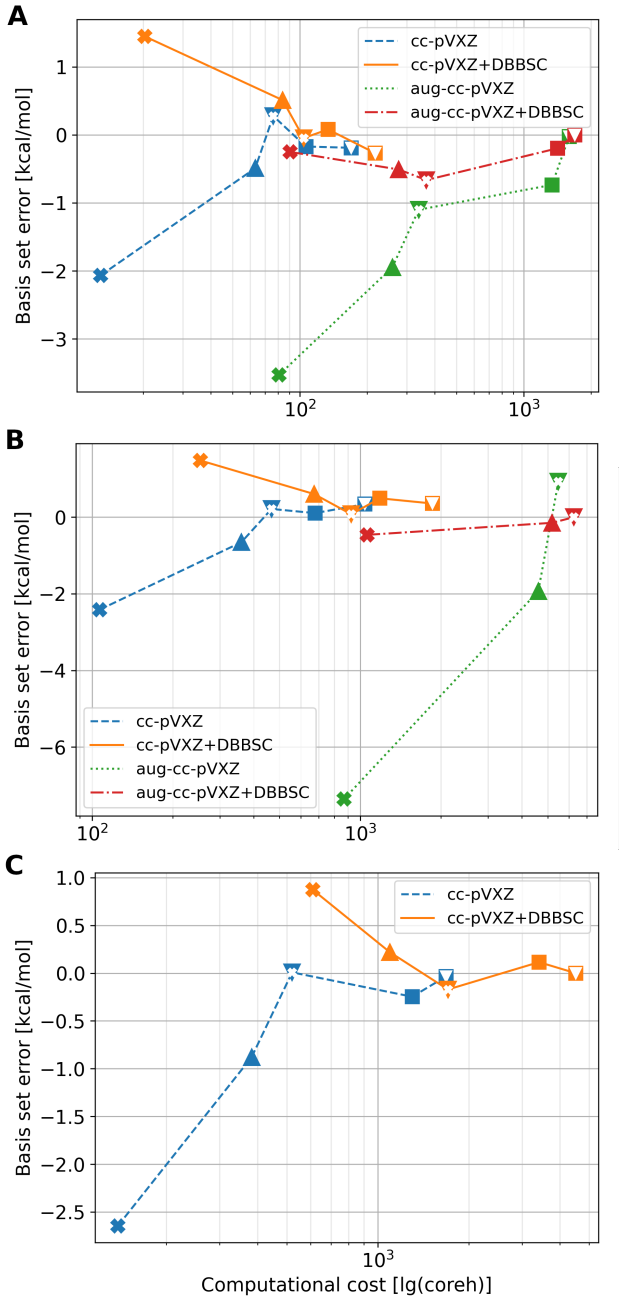

Figure 4: Convergence of the LNO-CCSD(T) first barrier (**TS1**–**RS**) with basis sets for three QM regions: **A**: 101, **B**: 238 and **C**: 372 atoms, as the function of the computational cost in core hours. Crosses, triangles and squares refer the cardinal numbers 2, 3 and 4 of the basis sets, combined to depict CBS extrapolation. The basis set errors are measured compared to the best converged CBS(T,Q) reference. DBBSC stands for the density-based basis set correction. Runtime for CBS data points are the sum of the corresponding two calculations. Basis set convergence for other stationary points are depicted in Figure S6.

QM atoms requires ca. 100 and 200 core hours for CBS(D,T) and CBS(T,Q), respectively (see Figure 4A). The cost increase is about 10-fold for the 238 QM atoms, but still only represent a few days using 16 cores. Then, the 372-atom LNO-CCSD(T) calculations took about 2000 core hours, indicating that computations approach the linear-scaling size range. Importantly, the QM-size-dependence for the LNO-CCSD(T)/CBS calculations is similar to that of the DFT relative values, which opens up the possibility to calculate the $\Delta\Delta E_{\mathrm{CC}}$ correction at a reduced cost. This concept is thoroughly discussed in section 4.2.

The convergence with local correlation thresholds is rapid, indicated by the LAF uncertainty measures[7,38] of the local approximation, defined as $\pm 0.5(E^X_{\mathrm{Tight}} - E^X_{\mathrm{Normal}})$. For the first barrier, this uncertainty is ±0.14 kcal/mol both with the 101 and 238 QM-atom regions with cc-pVTZ, which decreases to ±0.04 with aug-cc-pVTZ. The highest uncertainty is obtained for the 101-atom cc-pVTZ calculation of the **TS2**, still only ±0.22 kcal/mol, see more details in Table S1.

To continue with finding DFT models for $\mathrm{DFT_E}$ and $\mathrm{DFT_S}$ following our DFT error analysis protocol[8] we screened 19 generally well-performing, popular functionals (a few from each rungs from GGA to RSH, and the revDSD-PBEP86 double-hybrid) on the 101-atom QM region against LNO-CCSD(T)/CBS(aT,aQ) references (Figure 5A). These calculations were done with the cc-pVTZ basis set and dispersion correction. The detailed numerical results are collected in Table S2 of the SI. Semi-local functionals have a significant overstabilization of the $\mathrm{S_N}2$ type phosphate cleavage transition state (TS1). Among the most affordable GGA rung, PBE-D3 and PW91-D3 performs the best, but some of their errors are still 5–6 kcal/mol. The performance of hybrid and RSH functionals are similar to each other for all stationary points. Considering the errors on the difference between the two barriers, as a proxy for the quality of describing the chemistry, one of the best performing functional is PBE0-D3, (0.73 vs. the 1.66 kcal/mol of the LNO-CCSD(T)/CBS references in Table S2). Besides the DFT energies, we analyzed the density sensitivity of functional, in which PBE0, M06-2X and BHLYP do well among the hybrid rung (see Table S3 and details in section S4.1

of the SI).[8]

Then, for the QM regions where relative energies approach convergence, we also tested a few representative functionals (372 QM atoms, Figure 5B), to study the QM size dependence of DFT performance. The size-dependence of the DFT and reference calculations is similar (Table S4), and so is functional performance for all studied QM sizes. The above test suggests the choice of PBE0-D3 for this specific reaction.

Next, we examined the convergence with the basis sets used at the PBE0-D3 level. Using a double-ζ (DZ) basis is a simple way to limit the computational cost, and may be reliable in calculating relative energies if the error is consistent along the configurations. Scanning a series of basis sets up to quadrupole-ζ (QZ), we found the barrier to **TS1** underestimated with smaller basis sets, even when diffuse functions are added. Conversely, triple-ζ (TZ) and QZ sets show quick convergence and def2-TZVP is proven to be a sufficiently large set (cf. def2-SVP and def2-TZVP in Table S5). The DFT basis set convergence trends are consistent across all QM sizes studied (**TS1** barriers of ca. 19.3, 17.4 and 15.3 kcal/mol for 101, 238, and 372 atoms in Table S5, respectively). Most importantly, due to the inconsistent basis set error between the chemical steps, TZ accuracy is the minimum requirement for qualitatively accurate description of the reaction (Figure 5D). The presented convergence tests demonstrate that the PBE0-D3/def2-TZVP on a 372-atom QM region represent a great choice for DFT_E. Namely, this level is reasonably converged both in terms of accuracy (against the LNO-CCSD(T)/CBS references) and in QM size, close to the largest QM regions practically feasible (600–800 atoms).

Upon revisiting the convergence vs. the size of the QM region at DFT_E level (Figure 5C), results confirm the convergence at the 372-atom QM selection (rightmost green triangle in Figure 5C–D). The trends of the preliminary QM size screening are the same, selections that prioritize explicitly charged residues over radial extension are converging to the results with large QM sizes faster. Consequently, for the proposed computational scheme in Eq. 1, this screening identifies the functional (PBE0-D3), basis (def2-TZVP) and QM region size

(372 atoms) for the effective DFT_E method to approach the LNO-CCSD(T)/CBS values within about 1.5 kcal/mol (Table S4). This remaining deviation is very well corrected via the $\Delta\Delta E_{CC}$ term of Eq 1, see section 4.2.

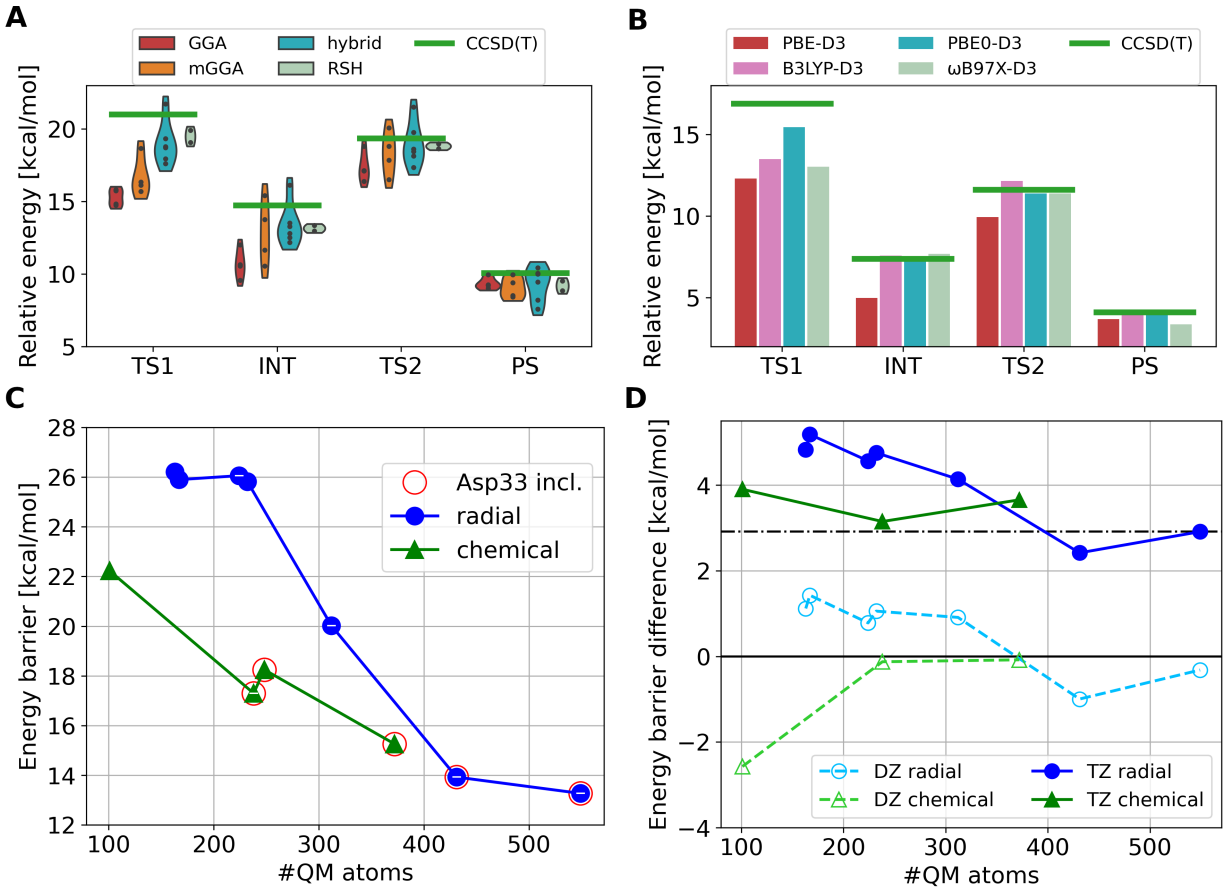

Figure 5: **A** Violin plot of DFT functional rungs compared to the LNO-CCSD(T)/CBS reference relative QM/MM energies, calculated for the 101-atom QM region. **B** Relative QM/MM energies calculated with selected DFT functionals vs. the LNO-CCSD(T)/CBS references for the 372-atom QM selection. **C** Convergence of the QM/MM energies (PBE0-D3/def2-TZVP/CHARMM36m) along two series of QM selections: blue circles: 'radial' extension from the reaction center, green triangles: selections prioritizing explicit charges, denoted as 'chemical'. Selections including Asp33 (and the complete network shown in 3**B**), are highlighted by red circles. Other relative energies are presented in Figure S8. **D** Difference in the barrier heights (TS1–TS2) calculated with PBE0-D3 and def2-SVP (DZ, dashed) and def2-TZVP (TZ, solid) basis sets.

As we set these parameters, we did not consider the possibility of running extensive free energy calculations beyond the fact that DFT_E gradients are readily available. When this

DFT_E model is too expensive for extensive sampling, one can look for compromises defining DFT_S for sampling extensively. The effect of small basis sets on the two transition states in this two-step reaction is quite unbalanced (DZ vs. TZ, 3–4 kcal/mol, Figure 5D and Table S5). On the other hand, shrinking the QM region from 372 to 238 atoms changes the overall barrier from 16.9 kcal/mol to 18.7 at the LNO-CCSD(T)/CBS level, while the DFT error and the relative position of the two TSs are virtually the same. Therefore it is more acceptable to make concessions in the size of the QM region than the size of the basis set. This is particularly important, as smaller basis set is often employed for optimization even if the final electronic energy is then corrected. With that in mind, the explicit-charge-specific selection of 238 QM atoms, at the same PBE0-D3/def2-TZVP level is a good candidate for DFT_S.

3.3 Three-layer QM_{HL}-in-QM_{LL}/MM models

As the QM atom lists and DFT parameters are now identified for DFT_E and DFT_S, we seek to retain their accuracy while accelerating their evaluation. Therefore we introduce a third layer in between the hybrid DFT and the MM regions. As identified above, the high-level, inner QM layer (QM_{HL}) is still treated at the PBE0-D3/def2-TZVP level, embedded into a more cost-effective QM_{LL} environment layer. In this section, we elaborate on the selection of the active atom list for the embedded QM subsystem and the level of theory for the outer QM_{LL} layer, corresponding to the total QM_{HL}+QM_{LL} region of 238 atoms.

We start with the Huzinaga-embedding and the selection of the QM_{LL} model. Similarly to the good performance of PBE0-D3 among hybrid functionals, PBE-D3 worked well at the GGA rung, so we selected it for QM_{LL} and screened the active atom selections. In general, it is good practice to select similar functionals for QM_{HL} and QM_{LL}.[52] As discussed in section 2.3, in our quantum embedding implementation, the starting point of the HL SCF calculation is created from the LL orbitals of the whole system. Consequently, the selection of the active subsystem in the Huzinaga embedding is more flexible and it does not require

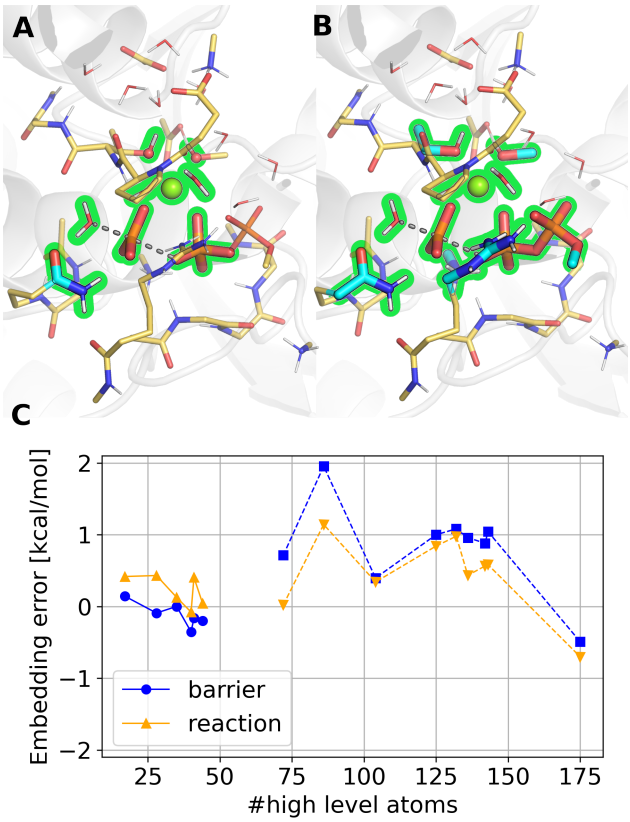

Figure 6: Active subset selection for quantum (**A** 28 atoms) and ONIOM embeddings (**B** 72 atoms). **C** Errors in relative energies against pure high level (PBE0-D3/def2-TZVP) calculations with different active atom lists in Huzinaga (solid lines) and ONIOM embedding (dashed lines). The environment in both cases was described at PBE-D3/def2-SVP level of theory. Errors depicted for the first (overall) barrier (TS1–RS) in blue (circles for Huzinaga quantum embedding, squares for ONIOM), and for the reaction energy (PS–RS) in yellow (upward triangles for Huzinaga quantum embedding, downward triangles for ONIOM).

link atom definitions. Conveniently, formal charges are not needed to individually assigned to all QM subsystems, like in ONIOM methods. An additional key difference compared to QM region selection (in QM/MM) is that the Huzinaga embedding interface can split polar bonds, even those with the bond order above one, or moderately delocalized moieties.[73] This enables smaller subsystems to be chosen for the more expensive QM_{HL} calculations. Initially, besides the reactive moieties, (the triphosphate anhydride, the glutamine (Q61) and the nucleophilic water), we added all of their counterions to the QM_{HL} atom list, namely a lysine (K16), an arginine (R789) and the Mg^{2+} with all of its coordinating ligands. This selection could be

farther shrunk by removing the lysine and arginine as well as the α-phosphate, arriving to 28 active atoms and 87 occupied MOs (Figure 6A). Interestingly, for these two elementary reaction steps, removing the Mg^{2+} ion with the alcohols and waters in its coordination sphere still works well (see 17 QM_{HL} atoms in Figure 6C). We anticipate this partitioning would break down if the Mg^{2+}-phosphate contact is broken, so for greater universality, we consider only the 28 active atom selection hereafter. Overall, the quantum embedded calculations reliably produce results within 0.5 kcal/mol of the complete QM_{HL} relative energies. This holds as long as all the reactive moieties are treated at the QM_{HL} level, which we also found in other types of reactions.[35] This consideration ensures that the moieties described with large semi-local DFT errors are treated at the hybrid QM_{HL} level, such as the S_N2 center of the phosphate cleaving **TS1**.

For an established 28-in-238 QM_{HL}-QM_{LL} split of the total QM region, we tested shrinking the AO basis set for the environment method. Despite the pronounced basis set error depicted in Figure 5D, this mixed basis approach (QM_{HL}: PBE0-D3/def2-TZVP in QM_{LL}: PBE-D3/def2-SVP) introduced negligible errors (< 0.4 kcal/mol, see Table S6) in the relative QM_{HL}-in-QM_{LL}/MM energies. The errors of this quantum embedded, mixed basis calculations (vs. the whole system treated at QM_{HL} level) are depicted in Figure 6C for active regions between 17 and 44 atoms. Employing the LESS framework to further accelerate the Huzinaga-embedded calculations introduces negligible deviations (0.2–0.3 kcal/mol), we discuss this further focusing on the achieved acceleration in section 4.1. For general cases, uncertainties related to the basis set truncation introduced in LESS are tightly and systematically controllable by the associated thresholds.[35]

To compare the performance of the Huzinaga quantum embedding to the most commonly used multi-layer scheme, ONIOM, we created a different series of subdivisions of the total QM region. Importantly, the selection rules for the active atoms have to be similar to the ones used at the QM/MM interface for keeping chemical integrity and for the definition of link atoms (hydrogens) in the HL ONIOM subsystem. This causes practical limitations opposed

to the flexible and effective region splitting enabled by the Huzinaga quantum embedding. Furthermore, unlike the Huzinaga embedding, the HL ONIOM calculations require the definition of charges for the LL atoms. Besides using the MM charges, they can be defined in multiple ways, for example, based on the LL QM calculations, as explored in section S5.2 in the SI. As a direct comparison to the quantum embedding, HL: PBE0-D3/def2-TZVP in LL: PBE-D3/def2-SVP ONIOM calculations were benchmarked with active atom lists between 72 and 175 atoms (Figure 6B). It is cautionary that the ONIOM relative energies do not show convergence below 100 HL atoms, and the ONIOM errors are twice as high as that of the Huzinaga quantum embedding, even when 75% of the total QM region (175 out of 238 atoms) is in the ONIOM HL layer (Figure 6C and Table S7). One advantage of the simple subtractive scheme of ONIOM is that even more cost-effective LL method can be used for the environment. For example, semi-empirical QM methods can be considered, such as tight binding DFT, that are not compatible with the Huzinaga embedding. We tested GFN2-xTB as an affordable tight binding method for the LL QM region, but the relative energies were insufficient for our ambitious accuracy purposes (Figure S9).

4 Discussion and General Suggestions

4.1 Accuracy as a Function of Computational Cost

We established that both DFT_E and DFT_S requires the accuracy of the hybrid/TZ level of DFT, while the difference between them is the size of the QM regions. We also demonstrated that this accuracy can be maintained with the application of mixed basis set and Huzinaga quantum embedding. The relative energies converge after 372 QM atoms, while at 238 atoms, the difference of the barrier heights provides reliable chemical conclusions. Now turning to the affordability aspect A3, Table 1 and Figure 7 summarize the computational cost of methods for both the 238- and 372-atom QM regions. The QM_{HL} method, PBE0-D3/def2-TZVP is selected as reference for the error in the barrier height, and speedup factors

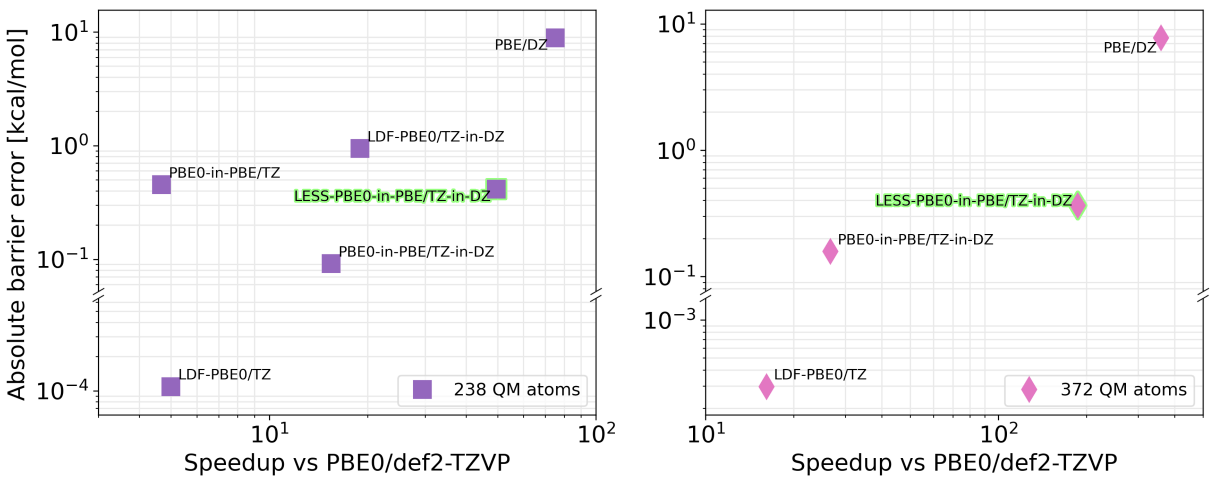

Figure 7: Absolute barrier error vs. speedup achieved against the approximation-free PBE0-D3/def2-TZVP calculations in kcal/mol for the 238 (left) and 372 (right) QM atom selections. DZ and TZ stand for def2-SVP and def2-TZVP basis sets, the -D3 labels are omitted for clarity.

are calculated against that same reference for both QM region sizes. For comparison, PBE-D3/def2-SVP results are also depicted (top right in Figure 7), which have a large error combined from the functional and the basis set (-8.86 and -7.72 kcal/mol), but are faster than PBE0-D3/def2-TZVP by 2–3 orders of magnitude.

Starting with the LDF-accelerated hybrid results, they agree with the conventional PBE0-D3 results within 10^{-3} kcal/mol, while being 5–12 times faster. Moreover, the speedup naturally increases with the system size (bottom left in Figure 7).

We tested a few additional combinations of approximations for the 238-atom QM regions. Our Huzinaga quantum embedding implementation without further approximations also fairs well for the barrier and it introduces a similar speedup to that of LDF (5.0 for LDF and 4.7 for Huzinaga). Introducing a smaller basis set to the $\mathrm{QM_{LL}}$ method (TZ-in-DZ mixed-basis) results in a 3-fold increase in speed for the quantum embedded calculations (from 4.7 to 15.5) without penalty on the accuracy. A similar mixed-basis approach combined with LDF, but without embedding, also comes with a similar speedup (from 5.0 to 18.9), although with a 0.94 kcal/mol error on the reaction barrier measured. The Huzinaga embedded calculations

with mixed-basis (-0.09 kcal/mol error, 15.5× speedup) come at a similar cost to the LDF PBE0/mixed basis runs, but with the accuracy closer to that of the slower PBE0-in-PBE/TZ or LDF PBE0/TZ methods.

Table 1: CPU time requirement and accuracy of DFT and DFT-in-DFT calculations.

method[a]	basis	approx.	QM atoms	CPU[b] time [coreh]	barrier error [kcal/mol]	speedup
PBE0	def2-TZVP	reference	238	138.7	-	1.0
PBE	def2-SVP		238	1.8	-8.86	75.0
PBE0	def2-TZVP	LDF	238	27.8	0.00	5.0
PBE0	TZ-in-DZ	LDF	238	7.3	-0.94	18.9
PBE0-in-PBE	def2-TZVP		28-in-238	29.6	0.45	4.7
PBE0-in-PBE	TZ-in-DZ		28-in-238	9.0	-0.09	15.5
PBE0-in-PBE	TZ-in-DZ	LESS	28-in-238	2.8	-0.41	49.5
PBE0	def2-TZVP	reference	372	1,045.1	-	1.0
PBE	def2-SVP		372	3.7	-7.72	281.6
PBE0	def2-TZVP	LDF	372	82.6	0.00	12.7
PBE0-in-PBE	TZ-in-DZ		28-in-372	50.1	0.16	20.9
PBE0-in-PBE	TZ-in-DZ	LESS	28-in-372	7.1	0.37	146.2

[a] all results include D3 empirical dispersion correction
[b] ran on 16 cores of an Intel Xeon Gold 6448H CPU without hyperthreading

The ultimate acceleration is, however, achieved by our LESS framework, which scales well with the system size, providing 50-fold and 146-fold speedups for the 238-atom and 372-atom QM regions, respectively. In these runs, the GGA level QM_{LL} environmental SCF computation take up more than half of the total execution time (cf. the PBE/def2-SVP speedups of 75 and 282 over PBE0/def2-TZVP). The LESS scheme, beyond the speedup of the approximation-free Huzinaga embedding, benefits from the effectiveness of mixed-basis approach and the truncation of the AO and DF basis sets in the QM_{HL} calculations. Furthermore, its efficiency is further enhanced by its compatibility with the in-core integral storage for the QM_{HL} calculation due to the LESS approximations.[35] Overall, the LESS QM_{HL}-in-QM_{LL} framework offers close to QM_{HL} accuracy for the cost of about two QM_{LL} calculations. In practical terms, if LESS is employed in the DFT_S model, a single point energy calculation takes 10 minutes wall time on 16 cores at an Intel Xeon Gold 6448H CPU

(2.8 core hour). This warrants the processing of tens of thousands of configurations within reasonable simulation times. DFT_E, which in this study differs only in the size of the QM region, would require about 3-times more resources per geometry, which makes exhaustive PES exploration sufficiently affordable.

The affordability of the herein reported QM models also make them good candidates for structure and path reoptimization, according to our preliminary calculations These will be thoroughly examined once the gradient calculations for LESS are available. All in all, the combination of DFT_S efficiency with CCSD(T)-corrected DFT_E energetics form a very good balance between the three aspects considered here, that are accuracy (A1), QM size (A2) and cost (A3).

4.2 Efficient Enthalpic Correction via LNO-CCSD(T)

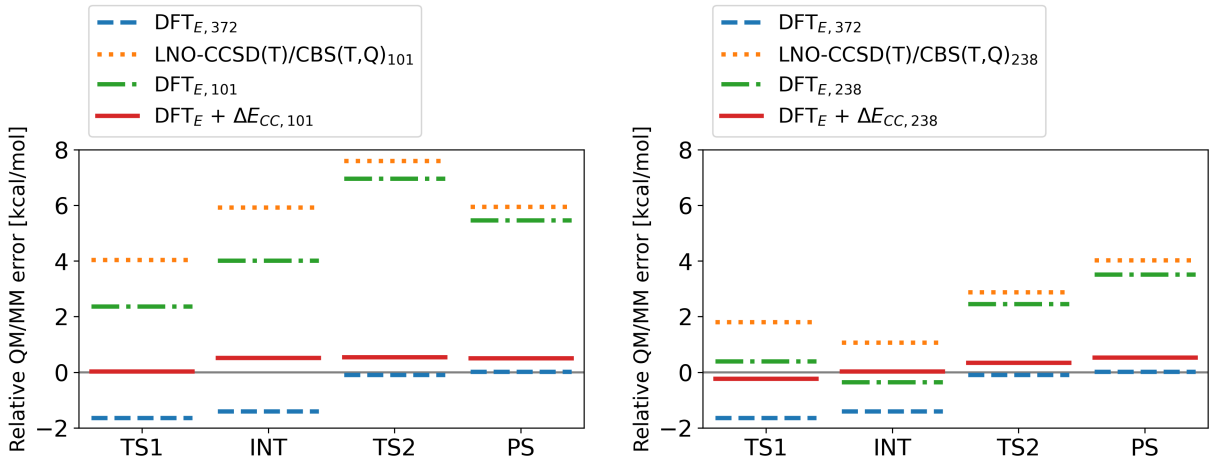

Figure 8: Errors in the relative QM/MM energies of stationary points using 101 (left) and 238 (right) QM atoms for the ΔE_{CC} correction, evaluated against the largest (372-atom) LNO-CCSD(T)/CBS reference values.

The LNO-CCSD(T) correction is used alongside DFT_E for a discretely represented reaction path. Here, we used it for a few stationary points and also characterize its cost to show its applicability for a larger number of conformers. The broader affordability enables the consideration of the LNO-CCSD(T) correction to improve the quality of the free energy

profile, e.g., via thermodynamic perturbation.[74,75] It is worth to note, that recent applications related to ML potentials highlight the difficulties to connect ensembles from different levels of theory, while also providing well-founded ways to overcome them.[76,77] While one may expect that the DFT_E or DFT_S methods selected to best reproduce the highly accurate references for the specific application provide sufficiently close ensembles, this assumption requires future validation.

For the largest, 372-atom QM regions, we carried out LNO-CCSD(T)/CBS calculations showing that convergence within chemical accuracy require CBS(D,T) basis extrapolation, taking more than 1000 core hours, while CBS(T,Q) results with DBBSC cost more than 4000 core hours. Figure 8 depicts that the well-performing DFT_E method has an uneven error along stationary points with respect to the LNO-CCSD(T)/CBS reference (blue dashes). When smaller QM regions are examined, the size-related error of the LNO-CCSD(T)/CBS and DFT_E calculations are similar (orange dots and green dot-dashes, respectively). In other words, the LNO-CCSD(T) correction term ΔE_{CC}, defined in Eq. 1 converges faster with the QM size than the relative energies and thus can be calculated at smaller QM sizes. In these CC-corrected composite references of $DFT_E + \Delta E_{CC}$, the relative barrier heights are at the quality of the LNO-CCSD(T)/CBS calculations, and the average error of the relative energies is 0.2–0.4 kcal/mol (Table 2). Crucially, the computational cost of these smaller LNO-CCSD(T)/CBS calculations with the 101-atom $\Delta E_{CC,101}$ correction are only 5% of that of the large, 372-atom model, ultimately allowing chemical accuracy achieved for 20-times more configurations. To put in perspective, the 217 core hour demand of the $\Delta E_{CC,101}$ calculations is in the same order as a hybrid DFT calculation of the 238-atom QM model without approximations.

4.3 General Advice on Model Building

The gain from the appropriate application of QM_{HL}-in-QM_{LL}/MM with negligible compromise in the high level method's accuracy provides an opportunity to elevate the overall

Table 2: Relative barrier heights, mean absolute errors [kcal/mol] and indicative computational cost [coreh] for LNO-CCSD(T)/CBS, DFT calculations and composite schemes.

method	QM atoms	ΔE(**TS1**–**TS2**) [kcal/mol]	MAE [kcal/mol]	CPU time [coreh]
LNO-CCSD(T)/CBS(T,Q)	372	5.21	-	4508
LNO-CCSD(T)/CBS(T,Q)	238	4.13	2.45	1856
LNO-CCSD(T)/CBS(T,Q)	101	1.66	5.88	217
DFT_E: LESS-PBE0-in-PBE/TZ-in-DZ	372	3.65	0.79	7.1
DFT_S: LESS-PBE0-in-PBE/TZ-in-DZ	238	3.15	1.68	2.8
LDF-PBE0-D3/def2-TZVP	101	0.62	4.70	1.0
$DFT_{E,372} + \Delta E_{CC,238}$		4.64	0.17	1866
$DFT_{E,372} + \Delta E_{CC,101}$		4.70	0.40	225

enzyme simulation practice. However, the introduction of the interfaces at QM/MM and QM_{HL}-QM_{LL} boundaries, selection of methods and atom lists for each layer add some complexity to the model definition. Therefore we compile the conclusions from our experience into transferable advice for setting up simulations for enzymatic reactions based on the introduced systematically converging strategies. Starting from an atomistic model (typically from experimental structures), we suggest the following pipeline:

1. Choose a reasonable, moderate sized QM region and DFT functional for preliminary calculations.

2. Obtain a preliminary reaction path (at least reactant, product and a few plausible intermediates and TSs) to represent the chemical space.

3. Establish the convergence of relative energies with the size of the QM region (A2 optimization, as proposed in section 2.3).

4. Select an appropriate DFT functional and basis set against reliable references (QM accuracy (A1) optimization).

   (a) Select an accurate high level (HL) DFT method and basis set for converged energetics [e.g. wrt. LNO-CCSD(T)].

(b) If warranted by QM region size, select an efficient low level (LL) DFT method and basis set for quantum embedding.

(c) Confirm size-convergence with the selected QM_{HL} approach. (A1+A2 optimization)

5. Optimize HL active atom list of the QM_{HL}-in-QM_{LL} embedding reproducing the QM_{HL} results.

   (a) If aiming also at sampling, select DFT_S e.g. further trim the total QM region if necessary. (A1+A2+A3 optimization)

6. Optimize/explore reaction paths (DFT_E) and proceed to free energy simulations (DFT_S).

   (a) If substantially different mechanisms are discovered, revisit steps 3–4.

7. Sample along key paths of *step 6* again and apply affordable ΔE_{CC} corrections via LNO-CCSD(T).

From a starting structure, typically representing the reactant or the product state of the reaction, one can define QM regions and start *step 3* converging reaction energies. However, we recommend the preliminary exploration of the chemical reaction(s) for two main reasons. First, it is important to include non-equilibrium geometries in the set of structures in which the convergence of the QM/MM model is tested. The overall reaction, thus which bonds are breaking and forming are known, so the a preliminary reaction channel can be created easily. The intermediate structures identified this way are expected to be more representative of the DFT uncertainties specific to the overall reaction than the reactant and product states alone (like the S_N2 TS in this example). Therefore including them enables more robust QM method selection for the bond rearrangements even if alternative paths are considered later. Second, depending on the reaction, the involvement and proximity of residues may change along the reaction coordinate, potentially resulting in an unbalanced representation in the QM region along the reaction coordinate when only the reactant is considered. This

way, important interactions along the path (like polar contacts presented here) can be better understood when determining the size of the QM region, that is used later for the mechanistic exploration. In the current work, the testing structures for *steps 1–2* were taken from a previous study.[20]

The general conclusion of *step 3* is that several hundred QM atoms may be required for converged properties. Our case study shows that the convergence is not necessarily monotonic, so accurate reference calculations and methods providing systematic convergence are extremely helpful for robust model selection. It is reasonable to assume that in the general case, a purely radial selection is suboptimal. Instead, following physical principles, such as prioritizing representation of explicit charges and polar contacts is expected to result in more effective QM region definition. Importantly, selection algorithms are more efficient, if they use fragments smaller than residues as the non-divisible unit, since most amino acid residues can be split along various apolar bonds at the QM/MM interface.

For the accuracy of the electronic structure method itself, gold-standard local-correlation methods, especially the highly optimized LNO-CCSD(T)[38] variant, are available for even the presumed several-hundred-atom QM region sizes, thus benchmarking production DFT functionals is clearly feasible (*step 4*). Studies indicate the need of hybrid or higher rungs of functionals, especially for accurate activation energies when it is helpful to suppress DFT errors, such as polarization and over-delocalization. When optimizing for speed, one might compromise on the size of the basis set, but caution is advised especially when different chemical steps are compared. Uneven error sources may change the conclusions qualitatively, like demonstrated on the basis set dependence of the relative barrier heights here. From the DFT benchmark one can also gain the information for an efficient QM_{LL} method used for the QM environment. Once the optimal functional-basis combination is found, the size-dependence studied in step 3 should be revisited.

Having the two-layer QM/MM baseline, in *step 5* the split of the QM regions by Huzinaga-based embedding can be intuitively constructed with a small active atom list (e.g. the reac-

tive atoms) without strict rules on the fragmentation. The interface does not require special definition, the only pivotal point is to make sure that the orbital selection is consistent along the reaction path. Robust orbital selection algorithms ensure this for sufficiently large QM_{HL} sizes beyond the region effected by the bond forming and breaking (although highly delocalized or metallic materials may be sensitive to this issue).[49] Besides the speedup quantum embedding offers, further techniques such as a mixed-basis definition for the QM subsystems, and ultimately the basis set truncation of our new LESS framework[35] can be greatly exploited. The individual approximations each constitute a 2–10-fold acceleration with well-controlled errors on the relative energies, making LESS QM_{HL}-in-QM_{LL} embedding up to two orders of magnitude faster than the conventionally required hybrid DFT methodology. In other words, in *step 5*, one establishes the DFT_E component of LASI, while in *step 5a* defines DFT_S. Concessions on the size of the QM region are justified if sampling methods necessitate them, especially if chemical conclusion remains intact and the introduced errors are well controlled by the underlying convergence study. In the presented case, the bespoke 238-atom region is deemed a reasonable compromise that can be used in subsequent free energy simulations, while the energetic term can be obtained with a better converged DFT_E approach.

Step 6 is where the (bio)chemical exploration takes place. Probing different mechanistic alternatives or free energy simulations can be executed with the selected LASI component, in an affordable and accurate fashion. If unforeseen reaction steps or intermediates are found with the more accurate model, in *step 6a*, one should verify the convergence and accuracy of the QM/MM model selected in *steps 3–5* on the new structures and path obtained with the optimized model. In other words, one should note the possibility that the initial model used to generate the initial structures and path in *steps 1–2* may substantially differ from the optimized model. In such cases, *step 6a* recommends to revisit *steps 3–5* on structures and path refined with the optimized method. The potential necessity and impact of this step are under investigation for the studied RAS reaction, which will be reported in our forthcoming

work. All in all, such extended mechanistic studies are enabled by two crucial criteria in our model selection: the feasibility of gradient calculations for the QM(-in-QM) model, enabling optimization and molecular dynamics; and restricting the calculation cost (i.e. wall time), ensuring practical affordability.

Finally, based on observations made in *step 4*, the application of ΔE_{CC} correction to DFT energetics may be very useful or even necessary to reliably approach chemical accuracy. Furthermore, one may expect that due to similar QM-size effects at the LNO-CCSD(T) and (well-selected) DFT levels, the incorporation of the ΔE_{CC} correction are sufficient using a smaller QM size, significantly decreasing the computational cost of the correction to the range of hundreds of core hours.

We recognize that not all simulations aim to calculate free energies. We summarize options the presented pipeline provides in Table 3. For potential energies, the $\mathrm{DFT_E}$ model should be effective, and then *step 5a* will not be necessary. If reliable benchmarks are available, one may choose to simplify *step 4* and/or *step 7*, but it is important to emphasize that simulation parameters can be coupled, and method (functional) selection should also be informed by QM size-dependence. Nevertheless, in chemical applications where the entire system may be described at the $\mathrm{QM_{HL}}$ level affordably, *steps 3* and *4c* are obsolete. While the number of demonstrations is limited, the generality of our systematic approach suggests that, a wide variety of applications would benefit from the acceleration that the novel LESS quantum embedding,[35] DBBSC[78] and LNO-CCSD(T)[7] schemes offer.

Table 3: Utility and performance of the components of the LASI thermodynamic scheme.

	$\mathrm{DFT_E}$	ΔE_{CC}	$\mathrm{DFT_S}$
goal	optimization, PES exploration	chemical accuracy	FES exploration
method	LESS hybrid-in-GGA	LNO-CCSD(T)/CBS	LESS hybrid-in-GGA
accuracy	1–1.5 kcal/mol	0.5–1 kcal/mol	$<$ 2 kcal/mol
QM size	300–400 atoms	$\approx$100 atoms	$\approx$200 atoms
comp. cost	10 coreh	200 coreh	3 coreh

5 Conclusions

This work presents a comprehensive strategy to advance the accuracy and efficiency of QM/MM simulations for enzymatic reactions, focusing on feasible computational cost while retaining high accuracy. Inspired by the its key characteristics of locality accelerated and systematically improvable, the new approach is labeled as LASI. Through systematic convergence studies on the QM region size and benchmarking against high-level local coupled-cluster reference calculations, our results echo that several hundred atoms could be required to achieve chemically accurate representations of enzymatic reaction energetics. Based on the presented strategy, a well-converged DFT_E methodology and a computationally lighter DFT_S model can be selected. DFT_E is ideal for the exploration of the potential energy surface, while DFT_S can be employed to efficiently explore reaction paths and perform free energy simulations, to obtain entropic contributions. Furthermore, we suggest a LNO-CCSD(T)-based a ΔE_{CC} correction to the relative electronic energies, and show that its application to DFT_E is effective and affordable, resulting in chemically accurate energetics even with 300–400 QM atoms. These levels of the computational scheme enables users to tailor their simulation protocol to available resources while confidently maintaining the best accuracy through systematic convergence and embedding strategies outlined in the study.

To make the DFT_E and DFT_S methods even more affordable, we employ our recently introduced LESS quantum embedding framework. It combines Huzinaga-equation-based embedding scheme with basis set truncation, accelerating the embedded hybrid DFT calculations by up to two orders of magnitude, while preserving the requisite precision of the high level calculation. This layered QM_{HL}-in-QM_{LL}/MM framework significantly surpasses the accuracy and efficiency of traditional ONIOM approaches, especially due to the flexible active atom selection and reduced basis sets in the environment. Importantly, we demonstrate that the method can handle complex enzymatic environments with highly polarized and charged active sites.

With the achieved speed-up, path optimizations and free energy simulations of (bio)-

chemical systems become affordable with the sufficiently converged LASI model. This overall aim of LASI simulations will soon be enabled by the development of analytical gradients to the LESS approach. Thus, its application refining the reaction path sampled here will be the focus of our forthcoming study, which will also reveal whether our optimized model has any dependence on our initial reaction path. Ultimately, we aim to move towards free energy simulations with the present biological system and beyond.

Our case study on Ras GTPase hydrolysis exemplifies the necessity of charge-aware QM region selection, surpassing simple radial extensions, and highlights that the nearby inclusion of polar contact networks is essential for convergence. The proposed workflow from preliminary path sampling through systematic convergence and embedding optimization to free energy simulations provides practical and transferable guidelines for setting up robust enzymatic reaction models in general. Overall, the new LASI framework paves the way for reliable and computationally feasible QM/MM simulations that can meaningfully complement experimental studies and aid the rational design of biochemical catalysts.

6 Computational Details

6.1 System and QM/MM Setup

As prepared in Ref. 20, the finite (non-periodic) QM/MM system (7815 atoms) was based on the PDB structure 1WQ1 containing the 5092 protein atoms and 906 water molecules environment. Electrostatic embedding and hydrogen link atom scheme was used were utilized in all calculations at the QM/MM interface, the MM terms for the link atoms were excluded from the energy evaluation. The MM terms were calculated by CHARMM 47b2, with the CHARMM36m and TIP3 force fields.[57,79–81] The non-bonding interaction energy MM terms were switched off in the range of 11–12 Å. For all calculations, we considered five stationary points identified in the previous works at the B3LYP/6-31+G* level of theory (Figure 1).

6.2 LNO-CCSD(T) Reference Calculations

LNO-CCSD(T) calculations employed the (aug-)cc-pVXZ basis sets (X = D, T and Q) on the first and second row elements and (aug-)cc-pV(X+d)Z for the phosphorus.[82] Weighted core-valence (aug-)cc-pwCVXZ[83,84] basis set were used for the Mg^{2+} ion to account for the sub-valence correlation of the 2s and 2p shells. We performed CBS extrapolations on the HF[85] and correlation[86] energies [CBS(X,X + 1)] to further accelerate the basis set convergence. Corrections from the complementary auxiliary basis set (CABS) to the HF, and the recent LNO-[39] and density-based basis set correction (DBBSC) to the CCSD(T) correlation energies were added.[78] LNO settings are used as defined in Ref. 38 and as implemented in the Mrcc program suite.[40–42]

6.3 DFT and DFT-in-DFT calculations

DFT calculations were carried out with Mrcc[40–42] with various functionals and the def2-X(Z)VP(PD) basis sets (X =S,T,Q).[87] Active MOs in the quantum embedding calculations were selected based on Mulliken population analysis with a threshold of 0.3. In LESS calculations, AO and DF basis set truncation were based on the Mulliken and LDF algorithms, with thresholds 10^{-4} and 2.0, respectively.[35] In all calculations, electrostatic embedding was used to best describe the environment and avoid issues with Kohn–Sham iterations.

6.4 Region Extension Algorithm for QM Region Definition

Our QM selection utility is available in the following repository:
`https://github.com/bertadenes/qm\_selector`. We considered all five stationary points when selecting QM regions. Upon the initial selection of key QM atoms, an extended core region of atoms with a given radius was defined, hydrogen atoms included. This region was iteratively extended along the covalent bonds defined in the MM topology, a link atom was initialized when a cutable bond was reached. Aliphatic fragments up to two carbon atoms

were excluded, if yielded by the selection. The produced inputs were processed by CHARMM to create the QM inputs with the correctly placed hydrogen link atoms.

Acknowledgement

The authors gratefully acknowledge support by ERC Starting Grant No. 101076972, "aCCuracy", János Bolyai Research Scholarship of the Hungarian Academy of Sciences, the National Research, Development and Innovation (NRDI) Office (Starting 149434, FK142489). The Hungarian Governmental Information-Technology Development Agency is acknowledged for awarding access to the Komondor and the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy) and the LEONARDO consortium.

Supporting Information Available

The Supporting Information is available free of charge at `url`. Detailed results of convergence with the size of the QM region, basis sets, local correlation thresholds; results of the functional benchmark and density sensitivity. Uncertainties related to embedding and ONIOM results. (PDF) Atom indices for entire QM and active subsystem selections. (XLS) Classical topology of the non-periodic MM system. (PSF) Cartesian coordinates of the studied stationary points. (CRD)

References

(1) Cui, Q.; Pal, T.; Xie, L. Biomolecular QM/MM Simulations: What Are Some of the "Burning Issues"? *The Journal of Physical Chemistry B* **2021**, *125*, 689–702.

(2) Nam, K.; Shao, Y.; Major, D. T.; Wolf-Watz, M. Perspectives on Computational Enzyme Modeling: From Mechanisms to Design and Drug Development. *ACS Omega* **2024**,

(3) Pöverlein, M. C.; Hulm, A.; Dietschreit, J. C. B.; Kussmann, J.; Ochsenfeld, C.; Kaila, V. R. I. QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling. *Journal of Chemical Theory and Computation* **2024**, *20*, 5751–5762.

(4) McFarlane, N. R.; Harvey, J. N. Minimum energy path methods and reactivity for enzyme reaction mechanisms: a perspective. *Pure and Applied Chemistry* **2025**, *97*, 1211–1222.

(5) Vennelakanti, V.; Nazemi, A.; Mehmood, R.; Steeves, A. H.; Kulik, H. J. Harder, better, faster, stronger: Large-scale QM and QM/MM for predictive modeling in enzymes and proteins. *Curr. Opin. Struct. Biol.* **2022**, *72*, 9.

(6) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479.

(7) Nagy, P. R. State-of-the-art local correlation methods enable accurate and affordable gold standard quantum chemistry up to a few hundred atoms. *Chem. Sci.* **2024**, *15*, 14556.

(8) Laczkó, G.; Pápai, I.; Nagy, P. R. Understanding DFT uncertainties for more reliable reactivity predictions by advancing the analysis of error sources. *J. Chem. Theory Comput.* **2025**, *21*, 9483–9497.

(9) Hu, L.; Söderhjelm, P.; Ryde, U. On the Convergence of QM/MM Energies. *Journal of Chemical Theory and Computation* **2011**, *7*, 761–777.

(10) Liao, R.; Thiel, W. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *Journal of Computational Chemistry* **2013**, *34*, 2389–2397.

(11) Jindal, G.; Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *The Journal of Physical Chemistry B* **2016**, *120*, 9913–9921.

(12) Roßbach, S.; Ochsenfeld, C. Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *Journal of Chemical Theory and Computation* **2017**, *13*, 1102–1107.

(13) Kulik, H. J. Large-scale QM/MM free energy simulations of enzyme catalysis reveal the influence of charge transfer. *Physical Chemistry Chemical Physics* **2018**, *20*, 20650–20660.

(14) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol *O*-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381.

(15) Cheng, Q.; DeYonker, N. J. The Glycine N-Methyltransferase Case Study: Another Challenge for QM-Cluster Models? *The Journal of Physical Chemistry B* **2023**, *127*, 9282–9294.

(16) Lu, X.; Fang, D.; Ito, S.; Okamoto, Y.; Ovchinnikov, V.; Cui, Q. QM/MM free energy simulations: recent progress and challenges. *Molecular Simulation* **2016**, *42*, 1056–1078.

(17) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living Journal of Computational Molecular Science* **2022**, *4*.

(18) McFarlane, N. R.; Harvey, J. N. Exploration of biochemical reactivity with a QM/MM growing string method. *Physical Chemistry Chemical Physics* **2024**, *26*, 5999–6007.

(19) Dürr, S. L.; Bohuszewicz, O.; Berta, D.; Suardiaz, R.; Jambrina, P. G.; Peter, C.;

Shao, Y.; Rosta, E. The Role of Conserved Residues in the DEDDh Motif: the Proton-Transfer Mechanism of HIV-1 RNase H. *ACS Catalysis* **2021**, *11*, 7915.

(20) Berta, D.; Gehrke, S.; Nyíri, K.; Vértessy, B. G.; Rosta, E. Mechanism–Based Redesign of GAP to Activate Oncogenic Ras. *J. Am. Chem. Soc.* **2023**, *145*, 20302.

(21) Raghavan, B.; Paulikat, M.; Ahmad, K.; Callea, L.; Rizzi, A.; Ippoliti, E.; Mandelli, D.; Bonati, L.; De Vivo, M.; Carloni, P. Drug Design in the Exascale Era: A Perspective from Massively Parallel QM/MM Simulations. *Journal of Chemical Information and Modeling* **2023**, *63*, 3647–3658.

(22) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *Journal of Chemical Theory and Computation* **2021**, *17*, 5745–5758.

(23) Schöller, A.; Kearns, F.; Woodcock, H. L.; Boresch, S. Optimizing the Calculation of Free Energy Differences in Nonequilibrium Work SQM/MM Switching Simulations. *The Journal of Physical Chemistry B* **2022**, *126*, 2798–2811.

(24) Snyder, R.; Kim, B.; Pan, X.; Shao, Y.; Pu, J. Bridging semiempirical and ab initio QM/MM potentials by Gaussian process regression and its sparse variants for free energy simulation. *The Journal of Chemical Physics* **2023**, *159*.

(25) Velázquez-Libera, J. L.; Recabarren, R.; Saez, D. A.; Castillo, C.; Ruiz-Pernía, J. J.; Tuñón, I.; Vöhringer-Martinez, E. Adapted DFTB3 Repulsive Potentials Reach DFT Accuracy for Hydride Transfer Reactions in Enzymes. *Journal of Computational Chemistry* **2025**, *46*.

(26) Song, L. F.; Rabara, D.; Bali, S. K.; Pei, J.; Lau, E. Y.; Kirshner, D.; Lightstone, F. C.; McCormick, F.; Stephen, A. G.; Yang, Y. How KRAS Mutations Impair Intrinsic GTP Hydrolysis: Experimental and Computational Investigations. *Journal of Chemical Information and Modeling* **2025**,

(27) Kolch, W.; Berta, D.; Rosta, E. Dynamic regulation of RAS and RAS signaling. *Biochemical Journal* **2023**, *480*, 1–23.

(28) Hobbs, G. A.; Der, C. J.; Rossman, K. L. RAS isoforms and mutations in cancer at a glance. *Journal of Cell Science* **2016**, 1287.

(29) Westheimer, F. H. Why Nature Chose Phosphates. *Science* **1987**, *235*, 1173–1178.

(30) Kamerlin, S. C. L.; Sharma, P. K.; Prasad, R. B.; Warshel, A. Why nature really chose phosphate. *Quarterly Reviews of Biophysics* **2013**, *46*, 1–132.

(31) Berta, D.; Buigues, P. J.; Badaoui, M.; Rosta, E. Cations in motion: QM/MM studies of the dynamic and electrostatic roles of H+ and Mg2+ ions in enzyme reactions. *Current Opinion in Structural Biology* **2020**, *61*, 198.

(32) Wineman-Fisher, V.; Delgado, J. M.; Nagy, P. R.; Jakobsson, E.; Pandit, S. A.; Varma, S. Transferable interactions of Li^+ and Mg^{2+} ions in polarizable models. *J. Chem. Phys.* **2020**, *153*, 104113.

(33) Maguire, M. E.; Cowan, J. A. Magnesium chemistry and biochemistry. *Biometals* **2002**, *15*, 203–210.

(34) Zhao, H.; Lőrincz, B.; Henkes, T.; Berta, D.; Nagy, P.; Tkatchenko, A.; Vuckovic, S. Accurate Density Functional Theory for Non-Covalent Interactions in Charged Systems. **2025**,

(35) Csóka, J.; Berta, D.; Nagy, P. Hybrid DFT quality thermochemistry and environment effects at GGA cost via local quantum embedding. *J. Chem. Theory Comput.* **2025**,

(36) Nagy, P. R.; Kállay, M. Optimization of the linear-scaling local natural orbital CCSD(T) method: Redundancy-free triples correction using Laplace transform. *J. Chem. Phys.* **2017**, *146*, 214106.

(37) Nagy, P. R.; Samu, G.; Kállay, M. Optimization of the linear-scaling local natural orbital CCSD(T) method: Improved algorithm and benchmark applications. *J. Chem. Theory Comput.* **2018**, *14*, 4193.

(38) Nagy, P. R.; Kállay, M. Approaching the basis set limit of CCSD(T) energies for large molecules with local natural orbital coupled-cluster methods. *J. Chem. Theory Comput.* **2019**, *15*, 5275.

(39) Mester, D.; Nagy, P. R.; Kállay, M. Basis-set limit CCSD(T) energies for large molecules with local natural orbitals and reduced-scaling basis-set corrections. *J. Chem. Theory Comput.* **2024**, *20*, 7453.

(40) Kállay, M.; Nagy, P. R.; Mester, D.; Rolik, Z.; Samu, G.; Csontos, J.; Csóka, J.; Szabó, P. B.; Gyevi-Nagy, L.; Hégely, B.; Ladjánszki, I.; Szegedy, L.; Ladóczki, B.; Petrov, K.; Farkas, M.; Mezei, P. D.; Ganyecz, Á. The MRCC program system: Accurate quantum chemistry from water to proteins. *J. Chem. Phys.* **2020**, *152*, 074107.

(41) Kállay, M.; Nagy, P. R.; Mester, D.; Gyevi-Nagy, L.; Csóka, J.; Szabó, P. B.; Rolik, Z.; Samu, G.; Csontos, J.; Hégely, B.; Ganyecz, Á.; Ladjánszki, I.; Szegedy, L.; Ladóczki, B.; Petrov, K.; Farkas, M.; Mezei, P. D.; Horváth, R. A. MRCC, *a quantum chemical program suite. See https://www.mrcc.hu/* **Accessed Dec 1, 2024**,

(42) Mester, D.; Nagy, P. R.; Csóka, J.; Gyevi-Nagy, L.; Szabó, P. B.; Horváth, R. A.; Petrov, K.; Hégely, B.; Ladóczki, B.; Samu, G.; Lőrincz, B. D.; Kállay, M. An overview of developments in the MRCC program system. *J. Phys. Chem. A* **2025**, *129*, 2086.

(43) Giner, E.; Pradines, B.; Ferté, A.; Assaraf, R.; Savin, A.; Toulouse, J. Curing basis-set convergence of wave-function theory using density-functional theory: A systematically improvable approach. *J. Chem. Phys.* **2018**, *149*, 194301.

(44) Loos, P.-F.; Pradines, B.; Scemama, A.; Toulouse, J.; Giner, E. A Density-Based Basis-Set Correction for Wave Function Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 2931.

(45) Bennie, S. J.; van der Kamp, M. W.; Pennifold, R. C. R.; Stella, M.; Manby, F. R.; Mulholland, A. J. A Projector-Embedding Approach for Multiscale Coupled-Cluster Calculations Applied to Citrate Synthase. *J. Chem. Theory Comput.* **2016**, *12*, 2689.

(46) Zhang, X.; Bennie, S. J.; van der Kamp, M. W.; Glowacki, D. R.; Manby, F. R.; Mulholland, A. J. Multiscale analysis of enantioselectivity in enzyme-catalysed 'lethal synthesis' using projector-based embedding. *Royal Soc. Open Sci* **2018**, *5*, 171390.

(47) Ranaghan, K. E.; Shchepanovska, D.; Bennie, S. J.; Lawan, N.; Macrae, S. J.; Zurek, J.; Manby, F. R.; Mulholland, A. J. Projector-Based Embedding Eliminates Density Functional Dependence for QM/MM Calculations of Reactions in Enzymes and Solution. *J. Chem. Inf. Model.* **2019**, *59*, 2063.

(48) Manby, F. R.; Stella, M.; Goodpaster, J. D.; Miller III, T. F. A Simple, Exact Density-Functional-Theory Embedding Scheme. *J. Chem. Theory Comput.* **2012**, *8*, 2564.

(49) Hégely, B.; Nagy, P. R.; Ferenczy, G. G.; Kállay, M. Exact density functional and wave function embedding schemes based on orbital localization. *J. Chem. Phys.* **2016**, *145*, 064107.

(50) Claudino, D.; Mayhall, N. J. Automatic Partition of Orbital Spaces Based on Singular Value Decomposition in the Context of Embedding Theories. *J. Chem. Theory Comput.* **2019**, *15*, 1053.

(51) Kolodzeiski, E.; Stein, C. J. Automated, Consistent, and Even-Handed Selection of Active Orbital Spaces for Quantum Embedding. *J. Chem. Theory Comput.* **2023**, *19*, 6643.

(52) Csóka, J.; Hégely, B.; Nagy, P. R.; Kállay, M. Development of analytic gradients for the Huzinaga quantum embedding method and its applications to large-scale hybrid and double hybrid DFT forces. *J. Chem. Phys.* **2024**, *160*, 124113.

(53) Nagy, P. R.; Samu, G.; Kállay, M. An integral-direct linear-scaling second-order Møller–Plesset approach. *J. Chem. Theory Comput.* **2016**, *12*, 4897.

(54) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. Fast Hartree–Fock theory using local fitting approximations. *Mol. Phys.* **2004**, *102*, 2311.

(55) Köppl, C.; Werner, H.-J. Parallel and Low-Order Scaling Implementation of Hartree–Fock Exchange Using Local Density Fitting. *J. Chem. Theory Comput.* **2016**, *12*, 3122.

(56) Zlobin, A.; Belyaeva, J.; Golovin, A. Challenges in Protein QM/MM Simulations with Intra-Backbone Link Atoms. *Journal of Chemical Information and Modeling* **2023**, *63*, 546–560.

(57) Hwang, W.; Austin, S. L.; Blondel, A.; Boittier, E. D.; Boresch, S.; Buck, M.; Buckner, J.; Caflisch, A.; Chang, H.-T.; Cheng, X.; Choi, Y. K.; Chu, J.-W.; Crowley, M. F.; Cui, Q.; Damjanovic, A.; Deng, Y.; Devereux, M.; Ding, X.; Feig, M. F.; Gao, J.; Glowacki, D. R.; Gonzales, J. E.; Hamaneh, M. B.; Harder, E. D.; Hayes, R. L.; Huang, J.; Huang, Y.; Hudson, P. S.; Im, W.; Islam, S. M.; Jiang, W.; Jones, M. R.; Käser, S.; Kearns, F. L.; Kern, N. R.; Klauda, J. B.; Lazaridis, T.; Lee, J.; Lemkul, J. A.; Liu, X.; Luo, Y.; MacKerell, A. D.; Major, D. T.; Meuwly, M.; Nam, K.; Nilsson, L.; Ovchinnikov, V.; Paci, E.; Park, S.; Pastor, R. W.; Pittman, A. R.; Post, C. B.; Prasad, S.; Pu, J.; Qi, Y.; Rathinavelan, T.; Roe, D. R.; Roux, B.; Rowley, C. N.; Shen, J.; Simmonett, A. C.; Sodt, A. J.; Töpfer, K.; Upadhyay, M.; van der Vaart, A.; Vazquez-Salazar, L. I.; Venable, R. M.; Warrensford, L. C.; Woodcock, H. L.; Wu, Y.; Brooks, C. L.; Brooks, B. R.; Karplus, M. CHARMM at 45: Enhancements in Accessibility, Functionality, and Speed. *The Journal of Physical Chemistry B* **2024**, *128*, 9976–10042.

(58) Clemente, C. M.; Capece, L.; Martí, M. A. Best Practices on QM/MM Simulations of Biological Systems. *J. Chem. Inf. Model.* **2023**, *63*, 2609.

(59) Sun, X.; Ryde, U. Reproducibility of QM/MM Calculations for the SARS-CoV-2 Main Protease. *Journal of Chemical Theory and Computation* **2025**, *21*, 7711–7723.

(60) Roca, M.; Maghsoud, Y.; Cisneros, G. A.; Świderek, K.; Moliner, V. Comparing Force Field Treatments in QM/MM Studies of the SARS-CoV-2 RNA-Dependent RNA Polymerase (RdRp) Mechanism. *Journal of Chemical Theory and Computation* **2025**,

(61) Karelina, M.; Kulik, H. J. Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *Journal of Chemical Theory and Computation* **2017**, *13*, 563–576.

(62) Brandt, F.; Jacob, C. R. Systematic QM Region Construction in QM/MM Calculations Based on Uncertainty Quantification. *Journal of Chemical Theory and Computation* **2022**, *18*, 2584–2596.

(63) Brunken, C.; Reiher, M. Automated Construction of Quantum–Classical Hybrid Models. *Journal of Chemical Theory and Computation* **2021**, *17*, 3797–3813.

(64) Csizi, K.; Reiher, M. Universal QM/MM approaches for general nanoscale applications. *WIREs Computational Molecular Science* **2023**, *13*, e1656.

(65) Wappett, D. A.; DeYonker, N. J. *Annual Reports in Computational Chemistry*; Elsevier, 2024; p 131–155.

(66) Wappett, D. A.; Cheng, Q.; Summers, T. J.; Santaloci, T. J.; Agbaglo, D. A.; Suhagia, T. A.; Banos, M. P.; DeYonker, N. J. RINRUS: A Toolkit for the Construction of Reproducible and Reliable QM-Cluster Models of Enzyme Active Sites. **2026**,

(67) Ho, J.; Yu, H.; Shao, Y.; Taylor, M.; Chen, J. How Accurate Are QM/MM Models? *The Journal of Physical Chemistry A* **2024**, *129*, 1517–1528.

(68) QM selector python utility by D. Berta at `https://github.com/bertadenes/qm_selector`.

(69) Grigorenko, B.; Polyakov, I.; Nemukhin, A. Mechanisms of ATP to cAMP Conversion Catalyzed by the Mammalian Adenylyl Cyclase: A Role of Magnesium Coordination Shells and Proton Wires. *The Journal of Physical Chemistry B* **2019**, *124*, 451–460.

(70) Ruiz-Pernía, J. J.; Świderek, K.; Bertran, J.; Moliner, V.; Tuñón, I. Electrostatics as a Guiding Principle in Understanding and Designing Enzymes. *Journal of Chemical Theory and Computation* **2024**, *20*, 1783–1795.

(71) Ganguly, A.; Boulanger, E.; Thiel, W. Importance of MM Polarization in QM/MM Studies of Enzymatic Reactions: Assessment of the QM/MM Drude Oscillator Model. *Journal of Chemical Theory and Computation* **2017**, *13*, 2954–2961.

(72) Böser, J.; Cui, Q.; Elstner, M.; Kubař, T.; Vuong, V.-Q. Non-covalent Interactions at the QM–MM Interface in the Semiempirical and Density Functional Limit. *Journal of Chemical Theory and Computation* **2025**, *22*, 678–695.

(73) Hégely, B.; Bogár, F.; Ferenczy, G. G.; Kállay, M. A QM/MM program for calculations with frozen localized orbitals based on the Huzinaga equation. *Theor. Chem. Acc.* **2015**, *134*, 132.

(74) Li, P.; Jia, X.; Pan, X.; Shao, Y.; Mei, Y. Accelerated Computation of Free Energy Profile at ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation. *Journal of Chemical Theory and Computation* **2018**, *14*, 5583–5596.

(75) Giese, T. J.; Zeng, J.; York, D. M. Multireference Generalization of the Weighted Thermodynamic Perturbation Method. *The Journal of Physical Chemistry A* **2022**, *126*, 8519–8533.

(76) Tkaczyk, S.; Karwounopoulos, J.; Schöller, A.; Woodcock, H. L.; Langer, T.; Boresch, S.; Wieder, M. Reweighting from Molecular Mechanics Force Fields to the

ANI-2x Neural Network Potential. *Journal of Chemical Theory and Computation* **2024**, *20*, 2719–2728.

(77) Giese, T. J.; Zeng, J.; York, D. M. Transferability of MACE Graph Neural Network for Range Corrected Δ-Machine Learning Potential QM/MM Applications. *The Journal of Physical Chemistry B* **2025**, *129*, 5477–5490.

(78) Mester, D.; Nagy, P. R.; Kállay, M. Basis-Set Limit CCSD(T) Energies for Large Molecules with Local Natural Orbitals and Reduced-Scaling Basis-Set Corrections. *Journal of Chemical Theory and Computation* **2024**, *20*, 7453–7468.

(79) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

(80) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(81) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

(82) Dunning Jr., T. H.; Peterson, K. A.; Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited. *J. Chem. Phys.* **2001**, *114*, 9244.

(83) Prascher, B. P.; Woon, D. E.; Peterson, K. A.; Dunning, T. H.; Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. VII. Valence, core-valence, and scalar relativistic basis sets for Li, Be, Na, and Mg. *Theor. Chem. Acc.* **2011**, *128*, 69.

(84) Hill, J. G.; Peterson, K. A. Gaussian basis sets for use in correlated molecular calculations. XI. Pseudopotential-based and all-electron relativistic basis sets for alkali metal (K-Fr) and alkaline earth (Ca-Ra) elements. *J. Chem. Phys.* **2017**, *147*, 244106.

(85) Karton, A.; Martin, J. M. L. Comment on: "Estimating the Hartree–Fock limit from finite basis set calculations". *Theor. Chem. Acc.* **2006**, *115*, 330.

(86) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639.

(87) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

Graphical TOC Entry