Biophysical Journal
Article

Biophysical Society

Cheminformatic quantum mechanical enzyme model design: A catechol-O-methyltransferase case study

Thomas J. Summers,[1] Qianyi Cheng,[1] Manuel A. Palma,[1] Diem-Trang Pham,[1,2] Dudley K. Kelso III,[1] Charles Edwin Webster,[3] and Nathan J. DeYonker[1,*]

[1]Department of Chemistry and [2]Department of Computer Science, The University of Memphis, Memphis, Tennessee; and [3]Department of Chemistry, Mississippi State University, Mississippi State, Mississippi

ABSTRACT To accurately simulate the inner workings of an enzyme active site with quantum mechanics (QM), not only must the reactive species be included in the model but also important surrounding residues, solvent, or coenzymes involved in crafting the microenvironment. Our lab has been developing the Residue Interaction Network Residue Selector (*RINRUS*) toolkit to utilize interatomic contact network information for automated, rational residue selection and QM-cluster model generation. Starting from an x-ray crystal structure of catechol-O-methyltransferase, *RINRUS* was used to construct a series of QM-cluster models. The reactant, product, and transition state of the methyl transfer reaction were computed for a total of 550 models, and the resulting free energies of activation and reaction were used to evaluate model convergence. *RINRUS*-designed models with only 200–300 atoms are shown to converge. *RINRUS* will serve as a cornerstone for improved and automated cheminformatics-based enzyme model design.

SIGNIFICANCE The efficiency, accuracy, and replicability of enzyme simulations is often hampered by ad hoc model design. To address this problem, we have developed the Residue Interaction Network Residue Selector (*RINRUS*) toolkit. *RINRUS* utilizes residue contact networks to automate construction of rational quantum mechanical cluster models. This work computes the reaction kinetics and thermodynamics for 550 *RINRUS*-designed models of the active site of catechol-O-methyltransferase, an enzyme that catalyzes the methyl transfer from S-adenosyl methionine cofactor to catechol substrates. Our results demonstrate the ability of *RINRUS* to rationally design small, reliable enzyme active site models and identifies chemical information useful for further model designs.

INTRODUCTION

For nearly two centuries, the structure, function, and catalytic power of enzymes have fascinated scientists, with countless studies seeking to understand their underlying mechanisms. Atomic-scale computer modeling of enzymes is currently a necessary part of the global multibillion-dollar research effort that aids the design of new pharmaceuticals, helps to investigate and engineer protein structure and function, and advances our understanding of the molecular basis of disease (1,2). The importance of atomic-level simulation of enzyme-catalyzed reactions was publicly acknowledged with the 2013 Chemistry Nobel Prize being awarded to Warshel, Levitt, and Karplus, who developed methods to treat the active site of an enzyme with quantum mechanics (QM) and the periphery with classical or molecular mechanics (MM) (3).

QM-only (also called QM-cluster), QM/MM, and our own N-layered integrated molecular orbital molecular mechanics (ONIOM) modeling are various approaches that have leveraged advancements in quantum mechanical theory and molecular dynamics to continually increase the ubiquity of computational enzymology (4–6). QM-cluster modeling has been shown to be particularly useful as a cost-effective method for studying structural and catalytic properties of enzyme active sites, especially those of metalloenzymes. As with all forms of modeling, the comparative accuracy of a model to reality is limited by the design of the model and relevant and reliable experimental data. For simulating the active site of enzymes, it is crucial to ensure

Submitted March 26, 2021, and accepted for publication July 29, 2021.

*Correspondence: ndyonker@memphis.edu

Editor: Diego Ferreiro.

https://doi.org/10.1016/j.bpj.2021.07.029

that not only the amino acids directly involved with the reaction are modeled at the QM level but also any residues, water molecules, ions, and coenzymes sterically and/or electrostatically crafting the active site microenvironment (4,7–9). Although this is a simple idea in principle, it is far harder in practice to identify rationally which residues must be partitioned into the QM level.

Although ad hoc protocols exist for selecting residues for inclusion in QM-level modeling, recommendations are typically ambiguous and generally inefficient (4,7). One of the most common practices is to simply include all residues within a certain radial distance from a point, perhaps the center of mass of substrate(s) or an active site metal center. Although suitable models could be constructed this way, calibration studies have confirmed large spheres (and consequently large models) are needed for the convergence of simulated enzyme thermodynamics and kinetics (8,10–18). These results are perhaps unsurprising, as nature does not enforce any geometric requirement to the design of an enzyme active site. Published "big-QM" models further add distant charged residues within the protein to generate 500–1000 atom models; however, the inclusion of less important residues unnecessarily increases the computational cost of any model (11,19,20). Attempts to quantify the importance of residues have been performed via a posteriori computations such as QM/MM thermodynamic cycle perturbations (21,22), linear response functions (23), or Fukui or charge shift analysis (14,24). However, such methods essentially require computational effort and thorough analysis of the constructed enzyme models to decide on an optimal model. Iterating an undirected residue selection process to self-consistency via QM or QM/MM computations is even more expensive.

Ideally, there would be a computationally inexpensive, a priori approach to enzyme model construction that utilizes structural and chemical data to rationally select residues (or parts of residues) for QM-cluster modeling. As a potential solution for this model creation problem, our lab has been developing the software Residue Interaction Network Residue Selector (*RINRUS*), which computes a contact-based residue interaction network (25,26) and uses the data to identify and rank residues for modeling. Further, *RINRUS* automatically trims and caps the residues via a rules-based criterion to form appropriate models and generates formatted input files for several popular electronic structure theory packages (see Materials and methods and Supporting materials and methods for details). The success of incorporating interaction and rules-based rationale into model design has been reported for QM-only models (27) and recently implemented into a QM/MM modeling application programming interface (28); however, there continues to be no definitive protocol for generalized QM-cluster enzyme model creation. Through establishing an automated and rigorous workflow, we envision solutions to several community-wide problems including standardization of enzyme QM-model creation, reducing learning curves for new users, and minimizing trial and error using poorly or incorrectly designed models. Implementing the *RINRUS* toolkit may also facilitate improving reproducibility of workflows and published results, a scientific community-wide need that has been most recently emphasized by the 2019 consensus study report *Reproducibility and Replicability in Science* released by The National Academies of Sciences, Engineering, and Medicine (29). To informally highlight the reproducibility problem within the QM/MM and QM-cluster modeling communities, we surveyed 58 QM/MM or QM-cluster model studies published within Jan 1–Mar 31 of 2015 and Jan 1–Mar 31, 2019, evaluating whether the models could be directly reproduced via reporting of Cartesian coordinates (see Supporting materials and methods for details). Only 20 studies (34%) reported Cartesian coordinates to the extent that reproduction is possible. Given the absence of consistent community reporting, embedding reproducibility via a systematic model design workflow would be a large step toward research standards in computational enzymology.

Ideally, the *RINRUS* workflow would be capable of identifying a singular or handful of models that best capture the balance between maximizing the number of key residues included to simulate the active site while minimizing the size of the QM region for computational efficiency. This leads to questions such as what makes the enzyme model "good"? What easily obtainable metrics might be universal in computational biochemistry for ranking the importance of interatomic and inter-residue interactions? We begin to answer these questions within the context of contact-based residue interaction networks (25,26).

The protein of interest for this case study is catechol-O-methyltransferase (COMT), a target enzyme of numerous QM-cluster and QM/MM studies (8,18,21,22,30–45). The mechanism catalyzed by COMT is rather simple, involving only an S_N2 methyl transfer from an *S*-adenosylmethionine (SAM) coenzyme to the oxygen of an Mg^{2+}-bound catecholate substrate (Fig. 1 *A*). Kinetic experiments on human COMT provide a free energy of activation ($\Delta G^{\ddagger}$) of 18–19 kcal/mol at 310 K (46,47), and computational studies report the methyl transfer reaction to be exergonic (8,34,35,43).

Previous computational studies have shown substantial variation in both $\Delta G^{\ddagger}$ and free energies of reaction (ΔG_{rxn}) with respect to QM-cluster or QM/MM model size. Recent results from QM/MM calibration studies using radial distance-based QM regions suggest that asymptotic convergence of thermodynamics and kinetics requires radial QM regions of 400–600 atoms (8,18,34). Unfortunately, conventional density functional theory (DFT) calculations of 400–600 atom models are prohibitively expensive for many research groups. The large QM region size required to study the COMT mechanism also defies conventional wisdom that kinetic and thermodynamic properties should converge quickly as the size of the QM region grows in a QM/MM partition. The slow convergence behavior of COMT has

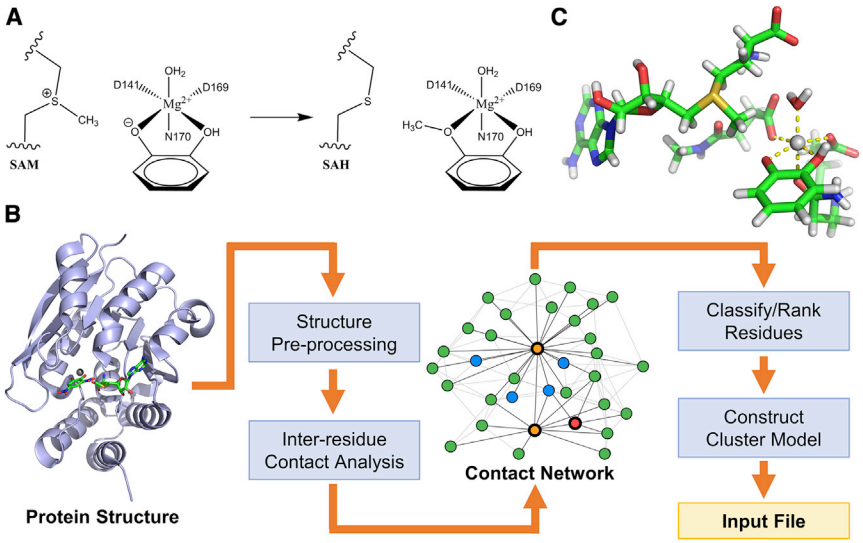

FIGURE 1 (*A*) COMT catalyzes the methyl transfer reaction from SAM to the oxygen of a Mg^{2+}-bound catecholate substrate, forming *S*-adenosylhomocysteine (SAH) and guaiacol. (*B*) The *RINRUS* workflow begins by processing a protein structure (x-ray, NMR, or computational simulation in PDB file format) before computing inter-residue contacts to form a contact network. Residues (*green*) and solvent (*blue*) interacting with the species of interest (the "seed," *orange* and *red*) are identified. Systematic classification or ranking schemes are used to construct appropriate cluster models. *RINRUS* then writes these models into an input file format appropriate for simulation in a variety of quantum chemistry software packages. (*C*) The base model from which all COMT models were built up. It is composed of the seed (SAM, CAT, Mg^{2+}), three residues, and one coordinating water completing the coordination of Mg^{2+} (D141, D169, N170, HOH411).

been attributed to the nonspherical active site, requiring an accurate description of both the Mg^{2+}-catechol coordination chemistry and the electrostatic stabilization of the large SAM cofactor (34).

Although the paradigm of calibrating expanding QM regions in a radial-distance-based fashion has been established to provide poor convergence for COMT, there is a surprising dearth of exploring alternatives to distance-based active site models in the literature. In this work, we present the reaction thermodynamics and free energies of activation for hundreds of QM-cluster models of COMT constructed by *RINRUS* using several possible workflows. By tracing the final results back to how the models were constructed, we seek to identify a construction protocol that consistently constructs accurate and efficient QM-cluster models of COMT. Though this work will only involve one case study, the findings from surveying an immense range of models of the same enzyme will allow future studies to invert the focus toward assessing the benefits of a particular approach on enzymes with more diverse structure and function. This cheminformatics perspective will be a rigorous step toward establishing a translatable, generalized computational enzymology protocol.

MATERIALS AND METHODS

The various structures and functions of proteins arise from the noncovalent interaction networks of their amino acid subunits. To highlight these networks, the complex three-dimensional structure of proteins may be simplified into a two-dimensional adjacency matrix or a graph mapping the residues to points (nodes) interconnected by lines (edges). Conventionally, each node represents an individual amino acid of the protein, and each edge represents a noncovalent interaction occurring between two amino acids. For more information on inter-residue contact networks and their design, properties, and applications within chemistry, the reader is directed to reviews by Giuliani (25) and Shen (48).

In this work, the construction of inter-residue contact networks begins by following a procedure similar to that of the software *RINerator* (26). First, hydrogens are added to the protein crystal structure (Protein Data Bank (PDB): 3BWM) using the program *Reduce* (49,50). As the 3BWM crystal structure has the inhibitor 3,5-dinitrocatechol coordinated to the active site metal, the two nitro groups were replaced with hydrogens to form the CAT substrate. An additional hydrogen was also added to the 2-amino functional group of the SAM substrate to bring it to a +1 charge, its expected protonation state. This modified crystal structure is the structure used for all subsequent network generation and model construction. The program *Probe* (51) is then used to identify noncovalent interactions throughout this structure. The program does this by rolling a small (0.25 Å radius) spherical probe over the van der Waals surface of the atoms and identifying both where the probe comes in contact with other noncovalently bound atoms and where van der Waals surfaces are clashing. The *Probe* output file details the contact or overlap "dots" for all of the atoms reflecting the distance of contacts or volume of overlaps. Wide contacts have an interatomic gap distance $\geq$0.25 Å, close contacts have an interatomic gap distance $<$0.25 Å, big overlaps have overlapping van der Waals radii $\geq$0.4 Å, small overlaps have overlapping van der Waals radii $<$0.4 Å, and hydrogen bonding is overlapping van der Waals radii between donor hydrogen and electronegative acceptor atoms (51). All of the reported contact dots (places where an interatomic contact or overlap occurs) are then collated for each residue to indicate which residues are interacting. The network illustrating all *Probe*-predicted contact interactions within 3BWM is shown in Fig. S1.

The chemically reactive species for this enzyme include the two substrates SAM and CAT along with the Mg^{2+} that CAT binds. One rationale for building up models of the active site would be to first focus on including residues immediately interacting with these reactive species. The network indicates this list includes 27 amino acids and four crystallographic waters. The specific parts of the residues having contact interactions with the reactive species (main chain or side chain) and the number of each contact type are provided in Table S1.

The base for building up all models described in this work is composed of the substrates SAM and CAT, Mg^{2+}, and the four species completing the coordination of Mg^{2+} (D141, D169, N170, and HOH411; Fig. 1). Residues are added to this base model by either assigning each residue an ordered rank or by adding groups of residues classified by a common feature. Models were automatically generated using the *RINRUS* software, trimming the models based upon a residue amino, carboxyl, or side chain having interatomic contacts with the seed. Places where covalent bonds are broken in trimming the model have hydrogens added to satisfy valency via the program *PyMol* v2.3.a0 (52). To maintain the general shape and semirigid character of the protein tertiary structure, all C_α atoms, along with the C_β atoms of Arg, Lys, Glu, Gln, Met, Trp, Tyr, and Phe side chains, were frozen to their crystallographic positions; research examining alternative ways of

treating these rigid atoms is underway (53,54). Further details about residue selection and model trimming are provided in the Supporting materials and methods. Although other research groups who employ QM-cluster models may have developed internal research protocols for trimming residues and fragments and freezing backbone atoms, we intend *RINRUS* to be the first, to our knowledge, enzyme model design toolkit to publicly codify these reproducible workflows (and also encourage hypothesis-driven testing of variations to our model-building decision trees).

All QM computations were performed using the Gaussian16 software package (55). The models were geometrically optimized using DFT with the hybrid B3LYP exchange-correlation functional (56,57). The computations used the 6-31G(d′) basis set for N, O, and S (58); the 6-31G basis set for C and H atoms (59); and the LANL2DZ effective core potential and basis set combination for Mg (60). The Grimme D3 (Becke-Johnson) dispersion correction was also included (61), along with a conductor-like polarizable continuum model using United Atom Topological Model sets of atomic radii, a nondefault electronic scaling factor of 1.2, and a dielectric constant of $\epsilon = 4$ (62,63). Unscaled harmonic vibrational frequency calculations were used to confirm all stationary points as either minima or transition states. Stationary points were found by first preoptimizing the model to the reactant structure. This preoptimized structure was then used to construct an approximate transition state structure by translating the methyl midway between the sulfur of SAM and the oxygen of CAT and flattening the methyl to a planar configuration. The transition state was optimized, and intrinsic reaction coordinate computations were used to confirm the formal reactant and product minima and calculate reaction free energies. Whether this procedure biases the simulated active site to more strongly stabilize the reactant structure (and whether such a bias would be of any significance) is unknown and an uninvestigated facet of computational enzymology.

The *k*-means clustering analysis (64) was run through *RStudio v.3.6.3* (65) using seed 3163 for replication purposes. Elbow and gap statistics (Fig. S6) were run using the *factoextra* package (66). For the gap statistic, the number of "bootstrap" Monte Carlo samples used was 50. Both elbow and gap statistics suggest using a k near $k = 6$ for the cluster analysis (Fig. S6). A $k = 6$ was ultimately used for further analysis because the clusters with $k = 6$ are reasonably partitioned into distinct groupings in which the range of free energies predicted by models within a cluster are not too broad (would happen with small *k*-clusters) and the interpretation of the clusters are not so narrow as to fail to be generalizable (would happen with large *k*-clusters). To identify the appropriate clusters, the Hartigan and Wong *k*-means clustering algorithm was used starting from a total of 50 different random starts (67).

RESULTS AND DISCUSSION

We began by computing a contact-based residue interaction network (Fig. 1 *B*) for an x-ray crystal structure of human COMT (PDB: 3BWM), in which residues, substrates, and solvent are illustrated as circles (termed "nodes" in standard graph theory nomenclature) interconnected by lines (termed "edges") when there are interatomic contacts between two residues or fragments. Although the construction and analysis of these graphs are already known to provide insight into allosteric regulation, protein folding and stability, and structure-function relationships (25,48), we repurpose the networks toward QM-cluster model design. The network indicated 27 protein residues and four crystallographic waters had contact interactions with any fragments central to the catalytic reaction (termed the "seed": SAM, CAT, or Mg^{2+}). The residue contacts with the seed were classified into five different types: wide contacts, close contacts, small overlaps, big overlaps, and hydrogen bonding. All QM-cluster models of COMT were constructed using the crystallographic coordinates of these residues and, unless otherwise indicated, trimmed according to the *RINRUS* workflow (refer to Supporting materials and methods). Models were expanded from the seed by one of two general ways: residues were incrementally added based upon a ranking criterion (e.g., distance from the seed or number of contacts with the seed) or groups of residues were added to the seed based upon similar residue features (e.g., type of interatomic contacts). The models constructed solely from the *RINRUS* contact information expand to a 485-atom model representing a "first interaction shell" maximal model that includes all residues with quantified contacts with any of the seed fragments. This maximal model is ellipsoidal in shape (Fig. 4 *B*), reflective of the nonspherical geometry of the COMT active site. Further details on the model-building schemes beyond what will be outlined in the discussion are provided in the Supporting materials and methods. In total, the methyl transfer transition state and connecting reactants and products for 550 unique QM-cluster models were computed. 1650 DFT-optimized stationary points were analyzed in this work.

Expansion of QM-cluster models by ranking of residues

We will first detail several ways COMT QM-cluster models were incrementally built up by ranking residues. The first metric is the current paradigm of ranking residues based on their distance to the active site. Though a simple distance metric may seem straightforward, this method can be ambiguous and tricky to replicate without reporting very precise definitions of the radial origin and the thresholds for adding residue fragments or entire residues. Subtle variances in definitions might qualitatively affect which residues or atoms are captured within varying radially expanding models. For this work, 25 models were constructed with *RINRUS* by incrementally adding residues ranked by the shortest distance from the position of any atom (including hydrogens) of the seed to the position of any atom of the surrounding residues. The models were expanded until all residues predicted by the contact network were incorporated, encompassing a 3.10 Å expansion from any atom of the seed. Two residues (K46 and N92) with no *RINRUS*-predicted contact interactions with the seed but that fall within the 3.10 Å distance threshold were necessarily included in these distance-based models.

Computed values of $\Delta G^{\ddagger}$ and ΔG_{rxn} are plotted against the distance-based expansion from the seed (Fig. 2 *A*). As the size of the model increases, the predicted $\Delta G^{\ddagger}$ converges (the $\Delta G^{\ddagger}$ is within ± 2 kcal/mol of the largest distance-based model) with QM-cluster models containing >342 atoms, but the predicted ΔG_{rxn} does not similarly converge even with the largest distance-based models. Some of the

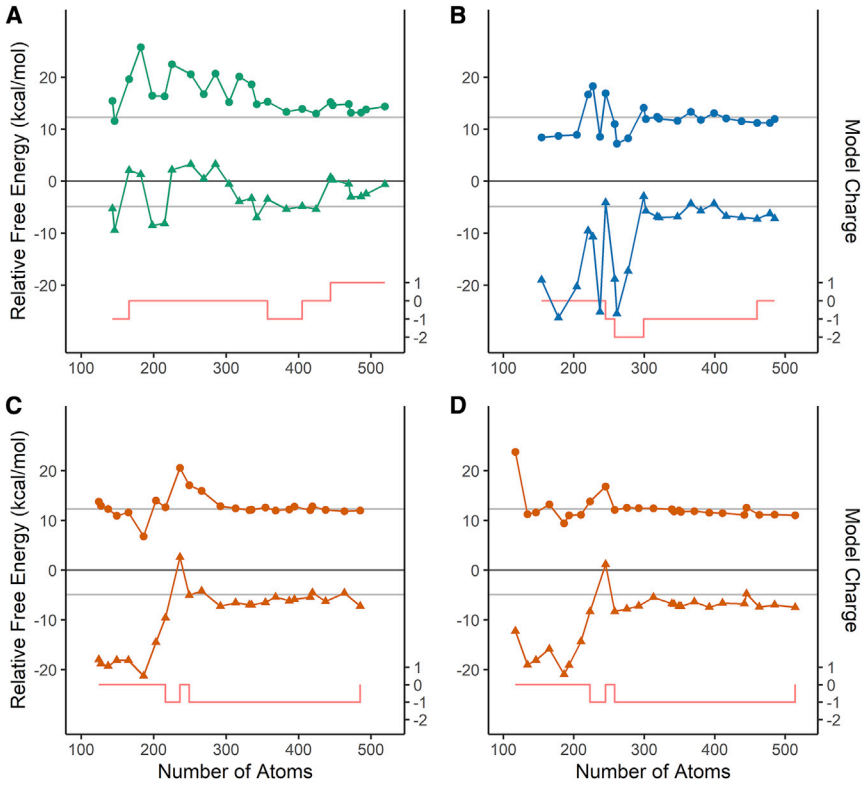

FIGURE 2 Computed methyl transfer $\Delta G^{\ddagger}$ (*circle*) and ΔG_{rxn} (*triangle*) free energies as models are systematically built up through different methods of ranking residues including distance from the seed (*A*), total number of contacts with the seed (*B*), frequency of residue in combinatoric scheme 2 sets (*C*), and a by-hand reconstruction of models by frequency of residue in combinatoric scheme 2 sets (*D*). Red lines indicate the charge for each model (*right axis*). Gray lines indicate the reference convergence values.

largest distance-based models computed in this work (containing 444 and 447 atoms) incorrectly predict an endergonic reaction.

The surprising appearance of qualitatively incorrect reaction free energies in the largest distance-based models brings up some crucial pitfalls in designing QM-cluster models but also ways that *RINRUS* can be used by the QM-cluster modeling community to circumvent these pitfalls. The convergence of the reaction free energy is disrupted by addition of the charged residue K46, which, as previously noted, does not have direct contact interactions with the seed. Such a qualitative shift in thermodynamic properties contradicts intuition that a larger QM-cluster model will always be "better" than a smaller model. At best, the addition of peripheral residues with no quantifiable interaction with seed residues or fragments adds unnecessary time to the DFT simulations, as observed with the addition of the uncharged N92 residue (not present in *RINRUS*-constructed models) changing $\Delta G^{\ddagger}$ and ΔG_{rxn} by <0.2 kcal/mol in the 486-atom distance-based model. It is known that balancing charged residues in models is important in COMT and other enzyme systems (4,8,11,20), but this COMT case study provides evidence that an undirected distance-based model-building scheme does not address this problem in a physically meaningful way. It may be fortuitous that the maximal COMT model generated by RINRUS (Fig. 4 *B*) does not include any boundary residues that are part of an unrequited charged pair, but even if it is coincidental, the *RINRUS* methodology provides a means for an automated solution to balancing charges. If the maximal model is thought of as a "first interaction shell" that encapsulates all residues that influence the active site chemistry, regardless of distance from the seed fragments, then the *RINRUS* source code can be easily adapted to include residues in the "second shell" that are necessary for charge balancing of larger-sized models. Testing of this procedure is currently underway by our lab.

As a step toward identifying a chemically directed way to expand models, we next considered the convergence of QM-cluster models constructed by ranking based on the number of contacts each residue has with the seed and incrementally building models from residues with the most contacts to fewest contacts with the seed. We define "convergence" in this study as being within ± 2 kcal/mol of the convergence reference values and remaining so as the model size is increased one residue at a time. The convergence reference values are defined as average relative free energies of the five largest models designed solely using *RINRUS* contact interactions: 12.3 kcal/mol for $\Delta G^{\ddagger}$ and -4.9 kcal/mol for ΔG_{rxn}. The converged reference value for ΔG_{rxn} is in agreement with other computational works reporting an exergonic reaction (8,34,35,43); $\Delta G^{\ddagger}$ is lower than the experimentally derived value, but this is expected considering the marginal level of theory used in this case study. The accuracy of *RINRUS*-derived models will be a subject of several future studies in our groups by varying level of theory, treatment of solvation, and approaches for freezing atoms, but for now, the consistency of the largest *RINRUS*

models provides a suitable reference point for convergence. With an improved ranking scheme using the number of residue-seed contacts, $\Delta G^{\ddagger}$ and ΔG_{rxn} both converge by the 302-atom model (Fig. 2 *B*). Although an interaction-based ranking fares better at prioritizing residues than distance-based expansion, there are some inherent limitations. Namely, larger residues with more surface area (e.g., lysine or tryptophan) are more likely to have more contacts with the seed and may bias the ranking compared to smaller residues. Ranking by number of contacts with the seed also does not weight or quantify the magnitude of electrostatic influences (e.g., charge, hydrogen bonding, and polarity). Nevertheless, even with this nonoptimal metric, constructing models by contact count still yields impressively small converged models.

Below, we will detail two combinatoric workflows for building models in which residues are classified into sets by common contact type. The third method for ranking residues involves ordering residues by the number of times each residue appears in a unique model from the combinatoric scheme 2 model sets (see below and Supporting materials and methods for details). This ranking inherently favors residues with more than one type of contact interaction. In using this residue ordering, $\Delta G^{\ddagger}$ and ΔG_{rxn} are converged when the QM-cluster model size is greater than ~300 atoms (Fig. 2 *C*), similar to the models designed through ranking residues by total contacts with the seed. The model with the greatest overestimation of $\Delta G^{\ddagger}$ and endergonic ΔG_{rxn} (236 atoms) corresponds to the addition of the positively charged residue, K144. The subsequent inclusion of the negatively charged E199 residue places the predicted free energies within qualitative accuracy, re-emphasizing the point that particular care in model design must be given toward charged residues and nearby residues that counter their effective charges.

Automation versus constructing QM-cluster models manually

The *RINRUS* package is still undergoing rapid development and needs further testing to address broader QM-cluster model design issues such as residue and substrate protonation states, the orientation of explicit solvent molecules, and conformational sampling (7,9). Although these factors may be manually addressed by the user, doing so places a potential bottleneck in the throughput of QM-cluster model applications.

In consideration of possible differences between manual and automated model building, models built by ranking residues via their frequency of appearance in combinatoric scheme 2 models (Fig. 2 *C*) were reconstructed by hand by the corresponding author (N.J.D.). The models were designed without any guidance from *RINRUS* beyond the identity of the specific residues in contact with the seed and their ranked order, and special attention was given toward residue protonation and sampling different conformations. The results of these "bespoke" models are presented in Fig. 2 *D* and are shown to be comparable to the models built by *RINRUS* (Fig. 2 *C*). There is reduced fluctuation in the $\Delta G^{\ddagger}$ for the smaller bespoke models versus comparably sized *RINRUS*-generated models, likely attributable to manual sampling of residue orientations, a treatment not done for any of the *RINRUS*-derived models. However, for the models greater than 300 atoms, there is no qualitative difference between the automated and the "by-hand" approach. These results demonstrate how *RINRUS*, even without carefully attending to residue protonation and conformational sampling, can construct QM-cluster models in a way similar to that by an experienced scientist but that is founded on a traceable cheminformatic basis and a reproducible, rational workflow. This automated efficiency will be important for future studies that may require constructing large numbers of models such as when sampling molecular dynamics simulations or exploring multistep chemical mechanisms.

Expansion of QM-cluster models by residue interaction features

The remaining models were built up from the seed by combining residues with common features, specifically by inter-residue contact type. The contact types contain two pieces of information used in QM-cluster model construction: the section of the residue contacting the seed (classified as residue main chain, residue side chain, or explicit water molecule) and the contact type (wide contact, close contact, small overlap, big overlap, or hydrogen bonding). Models were constructed by taking all combinations of the contact types and, for each combination, building a QM-cluster model using all residues with the specific contact types of that combination. These models represent a combinatoric approach to building up models by adding groups of residues by common features to the seed (combinatoric scheme 1; see Supporting materials and methods for details). To further increase the number of models and data set size, the sets of residues classified by contact types were repartitioned into a second combinatoric approach (combinatoric scheme 2; see Supporting materials and methods for details), although the generation of these sets is not rigorous or necessarily applicable to other biosystems. Given the limitations of time and resources, 114 (of 204 possible) models of combinatoric scheme 1 and 357 (of 736 possible) models of combinatoric scheme 2 have been simulated, representing all unique combination-based models up to at least 320 atoms (Fig. S5). As the goal is identifying small yet accurate QM-cluster models, the cost of expanding the data set to include hundreds of additional large models is not expected to lead to substantial improvements in analysis.

In plotting $\Delta G^{\ddagger}$ and ΔG_{rxn} of QM-cluster models built through the two combinatoric schemes (Fig. 3, *A* and *B*), a

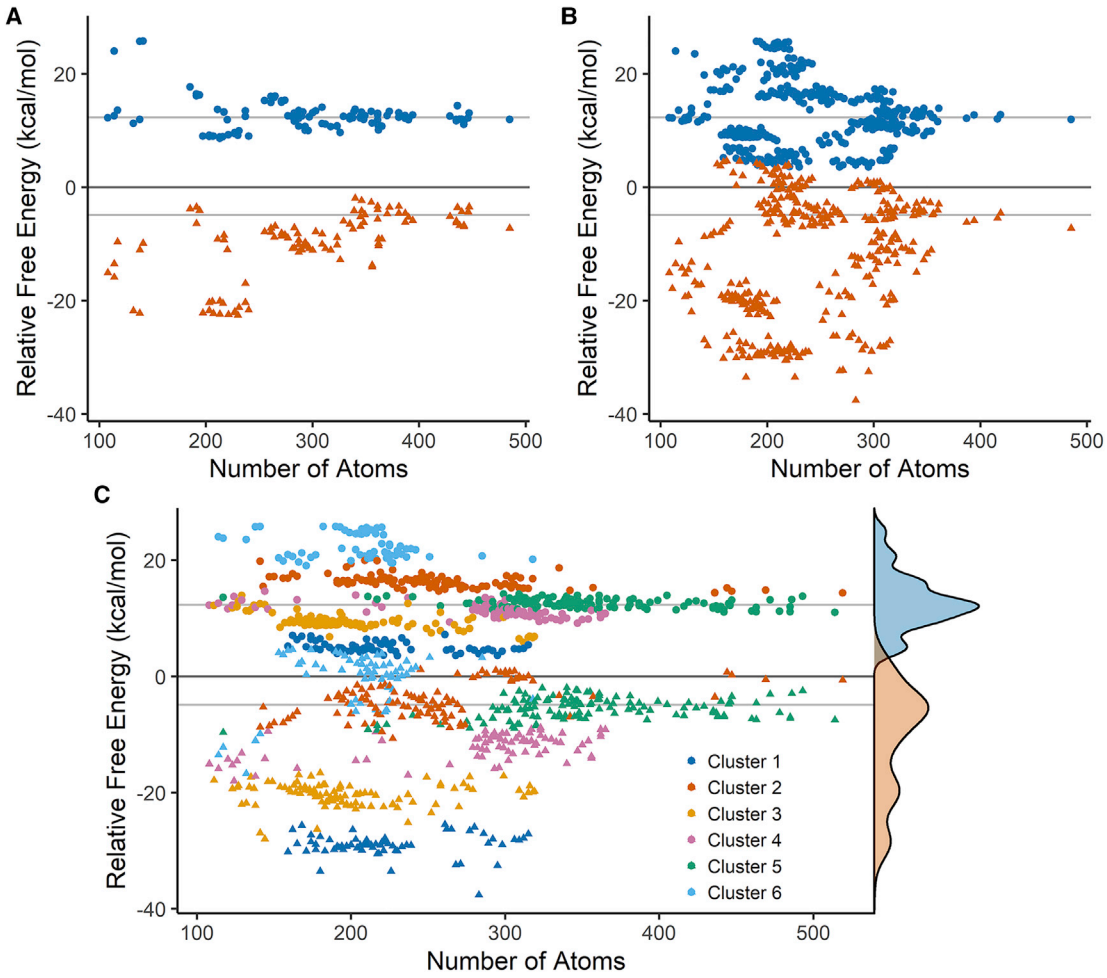

FIGURE 3 Computed methyl transfer $\Delta G^{\ddagger}$ (*circle*) and ΔG_{rxn} (*triangle*) as models are constructed through either the combinatoric scheme 1 (*A*) and combinatoric scheme 2 (*B*). (*C*) Scatter and density plot of $\Delta G^{\ddagger}$ (*blue* density) and ΔG_{rxn} (*tan* density) for all simulated models. Six clusters identified by *k*-means clustering of similar $\Delta G^{\ddagger}$ and ΔG_{rxn} are differentially colored. Gray lines indicate the reference convergence values.

wide range of computed kinetic and thermodynamic values were exhibited. Variation in $\Delta G^{\ddagger}$ and ΔG_{rxn} originates from differences in model composition rather than models optimizing into unnatural orientations, as the root mean-square deviation of unconstrained residue heavy atoms of the geometry optimized reactant state compared to the x-ray crystal structure is, on average, only 0.53 Å for all models (Fig. S4; standard deviation, 0.17 Å). Similar to the models built by ranking residues, there are models with fewer than 300 atoms that yield accurate predictions, affirming that QM-cluster model convergence for COMT does not require >400 atom models.

Identifying important residues

A general grouping of COMT QM-cluster models that predict similar (though not necessarily accurate) free energies is observed in Fig. 3 for both combinatoric schemes. This leads to the question of which residues are required to form an accurate model. To more clearly distinguish the grouping of unique models that predict similar kinetic and thermodynamic properties, the *k*-means clustering algorithm was used to partition the entire data set of unique QM-cluster models into six groups (Fig. 3 *C*) based upon their predicted $\Delta G^{\ddagger}$ and ΔG_{rxn} (64). Though an unsupervised method was used to group the models, the identified clusters are reasonable and properly differentiate the models with both converged $\Delta G^{\ddagger}$ and ΔG_{rxn} (cluster 5) from markedly inaccurate models (clusters 1 and 6), as well as models with converged values for either $\Delta G^{\ddagger}$ or ΔG_{rxn}, but not both (clusters 2, 3, and 4).

The residues that differ among the clusters give insight into which residues have a comparably strong influence on convergence. Tabulating the percent occurrence of each residue within the COMT models of each cluster (Figs. 4 and S7; Table S2), nine residues present in >90% of the cluster 5 models are absent or have a greatly reduced presence in other clusters. For example, in the models of cluster 6, which systematically overestimate $\Delta G^{\ddagger}$ and 65% of which incorrectly predict an endergonic reaction, none contain

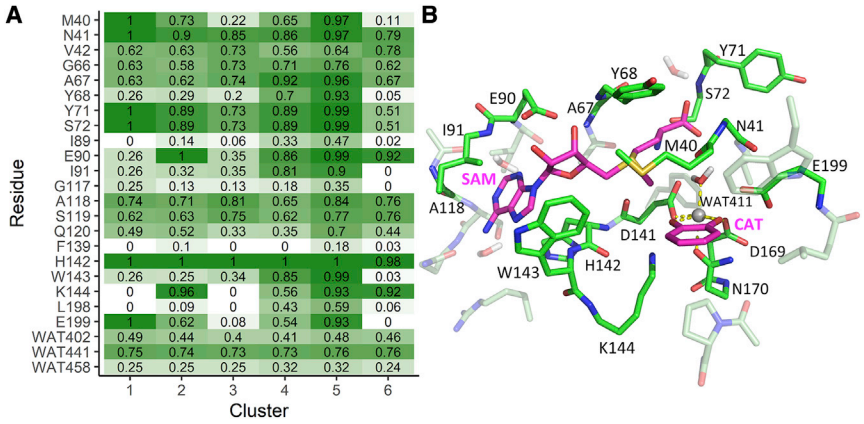

FIGURE 4 (*A*) Relative frequency for each residue being present in the models of a *k*-cluster. Values are proportionally shaded to emphasize differences in residue composition among *k*-clusters. (*B*) Visualization of the maximal 485-atom model highlighting the residues that occur in >80% of cluster 5 models. The carbon atoms of the substrates are colored magenta.

E199 and only 11% contain M40. Without these residues, the QM-cluster models are missing 1) the stabilizing hydrogen bonding interactions between E199 and CAT and 2) the hydrophobic interactions between M40 and the SAM, resulting in consistently large deviations with respect to the converged free energies.

Surprisingly, residues identified as particularly important for convergence are not always localized around the atoms directly involved in the methyl transfer. For instance, E90 (which is present in 99% of the models in cluster 5 but only in <35% of the models in clusters 1 and 3) is ~10 Å from CAT but plays a role in stabilizing and properly orienting the SAM. Other residues such as I91, A118, S119, and H142 are present in >70% of the models in cluster 5 and appear to play important roles in crafting the active site microenvironment.

With residues crucial for accurate QM-cluster modeling of COMT identified, the next step is to examine contact and classification metrics to see whether any were particularly suitable for predicting the relative importance of residues. For the contact classifications, there is unfortunately no consistent combination of contact types among the cluster 5 models for yielding converged models. Using the total contacts between the seed and each residue (Fig. 2 *B*) as a ranking system proves modestly successful, as 9 of the 13 residues present in >80% of the cluster 5 models have a high frequency of contacts with the seed and would be correctly prioritized. The four residues with low contacts (N41, A67, Y71, and A118) are adjacent to high-contact residues and largely have main chain interactions with the seed, explaining the fewer contacts. The general success of using total contacts as a ranking scheme was previously shown in Fig. 2 *B*, in which converged models had 302 atoms as a lower bound. Improvements to this ranking method are warranted (and are under current investigation by our lab), ranging from incorporating additional chemical descriptors for the interatomic contacts (e.g., through *Arpeggio* (68)) to developing a weighting system to favor certain contact interactions (e.g., hydrogen bonding, polar, or aromatic). In the end, *RINRUS* provides a computationally inexpensive, rational, and reproducible means to building enzyme QM-cluster models.

CONCLUSIONS

Computational enzymology has made incredible impacts on understanding the atomic-level intricacies of enzyme function. Although computational resources and scaling limitations of quantum chemistry are among factors limiting progress in this field, little attention has been given toward how poor or irreproducible model design might be hampering scientific progress. Many publication-quality enzyme models have been founded on rationale not necessarily suited for modeling nonspherical active sites (e.g., radial distance criterion) or via rationale prone to fallibility (a researcher's chemical intuition). Techniques addressing this problem by identifying important residues a posteriori have been useful but fail to meet the need for a computationally inexpensive a priori method for designing enzyme models.

As a step toward addressing community-wide problems in computational enzymology, we have been developing the *RINRUS* toolkit to automate the residue selection and construction of QM-cluster models. *RINRUS* utilizes the cheminformatics of interatomic contact networks as the rationale for identifying active site residues and ranking and classifying them. The catalytic methyl transfer reaction of the human COMT enzyme was simulated with a total of 550 unique models, illustrating how information from *RINRUS* was used to build models up from a base structure by either adding residues incrementally via a ranking scheme (e.g., total contacts with the seed) or by adding combinations of groups of residues (e.g., type of contacts). Clusters of models with common predictions of reaction and transition state free energies were compared to identify residues important for accurate simulations of COMT. Tracing the converged models and important residues back to how the models were constructed revealed that ranking residues by the frequency of their contacts with the seed was a particularly useful method, with QM-cluster models with 210–300 atoms yielding converged thermodynamic and kinetic properties. Additionally, the

methodology employed by *RINRUS* to identify seed-residue interactions and accordingly trim QM-cluster models favorably compares to that of "by-hand" models created by an experienced computational biochemist.

The major focus of this work has been to quickly converge energetic properties of smaller QM-cluster models to those of a maximally sized QM-cluster model. Further testing of the QM-cluster modeling methodology for accuracy to other well-defined experimentally known quantities (e.g., NMR chemical shifts) is an obvious next step for our lab to take. However, proper calibration of QM-based computational enzymology is contingent upon first developing a rational and reproducible scheme for building QM-cluster models. Particular avenues of study include calibration of DFT, one-electron basis set, implicit solvation parameters, empirical dispersion corrections, and other variables of electronic structure theory to truly assess the accuracy of QM-cluster modeling beyond a metric of internal consistency. Recent developments in linear scaling coupled cluster theory suggest ways to incorporate more rigorous "black box" electronic structure theories into the realm of computational enzymology. Investigating the structural and cheminformatic variation from constructing models using x-ray crystal structures versus conformational sampling frames from molecular dynamics simulations is also underway. These studies are in concert with investigations by our lab on improving the chemical descriptors and ranking schemes, integrating machine learning into the workflow, and examining how to best account for the impact that charged residues have on modeling the active site. In the future, we also seek to expand the functionality into automating QM/MM modeling construction. A forthcoming publication will describe the *RINRUS* software package and include thorough tutorials. Public availability and adoption of *RINRUS* will substantially reducing learning curves for new practitioners of QM-cluster modeling and initiate a feedback loop for improving the generalizability of *RINRUS* for constructing QM models of proteins beyond COMT and the enzymes studied within our lab.

Though model design and reproducibility questions have been largely ignored within the greater computational enzymology community, we hope this work will foster self-reflection on the underlying assumptions behind how atomic-level enzyme simulations are derived. The current practices often require unnecessarily large models to obtain accurate or internally converged results, which is limiting progress and is undoubtedly daunting to inexperienced chemists and biochemists interested in contributing to the field. Through the automated workflows provided by *RINRUS* and its successful results demonstrated in this work, we present the first steps, to our knowledge, toward discovering and implementing a computationally inexpensive, cheminformatic-based means for constructing reproducible, rational, and rigorous enzyme models. Admittedly, this case study of a single enzyme does not fully address all parameters of QM-cluster enzyme model construction. Nevertheless, reproducible workflows in computational enzymology, supported by *RINRUS* development, will improve openness and data sharing and facilitate novel cyber- and software infrastructure in biochemistry and biology.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.07.029.

AUTHOR CONTRIBUTIONS

T.J.S., Q.C., and N.J.D. designed research. T.J.S., M.A.P., Q.C., and N.J.D. performed research. T.J.S., M.A.P., Q.C., D.-T.P., and D.K.K. wrote code. D.K.K., C.E.W., and N.J.D. developed the initial concept and workflows. T.J.S., M.A.P., Q.C., C.E.W., and N.J.D. analyzed data. T.J.S., Q.C., C.E.W., and N.J.D. wrote the manuscript.

ACKNOWLEDGMENTS

The authors thank Professor Ramin Homayouni (Oakland University) for helpful discussions with this work.

This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under grant number 1451514 (for T.J.S.), NSF CAREER BIO-1846408 (for N.J.D.), and NSF CAREER CHE-0955723 and CHE-1543490 (for C.E.W.). This work was also supported by start-up funding from the University of Memphis Department of Chemistry and in part by a grant from the University of Memphis College of Arts and Sciences Research Grant Fund (D.-T.P. and N.J.D.). The High Performance Computing Center and the Computational Research on Materials Institute at The University of Memphis also provided generous resources for this research.

SUPPORTING CITATIONS

References (69,70) appear in the Supporting material.

REFERENCES

1. Kiss, G., N. Çelebi-Ölçüm, …, K. N. Houk. 2013. Computational enzyme design. *Angew. Chem. Int. Ed. Engl.* 52:5700–5725.
2. Kollman, P. A., B. Kuhn, and M. Peräkylä. 2002. Computational studies of enzyme-catalyzed reactions: where are we in predicting mechanisms and in understanding the nature of enzyme catalysis? *J. Phys. Chem. B.* 106:1537–1542.
3. The Nobel Foundation 2013. The Nobel Prize in Chemistry. 09 June 2020 https://www.nobelprize.org/prizes/chemistry/2013/summary.
4. Ahmadi, S., L. Barrios Herrera, …, D. R. Salahub. 2018. Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: a tutorial review. *Int. J. Quantum Chem.* 118:e25558.
5. Kmita, K., C. Wirth, …, V. Zickermann. 2015. Accessory NUMM (NDUFS6) subunit harbors a Zn-binding site and is essential for biogenesis of mitochondrial complex I. *Proc. Natl. Acad. Sci. USA.* 112:5685–5690.
6. Li, X., P. E. M. Siegbahn, and U. Ryde. 2015. Simulation of the isotropic EXAFS spectra for the S2 and S3 structures of the oxygen evolving complex in photosystem II. *Proc. Natl. Acad. Sci. USA.* 112:3979–3984.

7. Lonsdale, R., J. N. Harvey, and A. J. Mulholland. 2012. A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* 41:3025–3038.
8. Kulik, H. J., J. Zhang, …, T. J. Martínez. 2016. How large should the QM region be in QM/MM calculations? the case of catechol O-methyltransferase. *J. Phys. Chem. B.* 120:11381–11394.
9. Borowski, T., M. Quesne, and M. Szaleniec. 2015. QM and QM/MM methods compared: Case studies on reaction mechanisms of metalloenzymes. *In* Combined Quantum Mechanical and Molecular Mechanical Modelling of Biomolecular Interactions, Advances in Protein Chemistry and Structural Biology T. Karabencheva-Christova, ed. . Academic Press Inc., pp. 187–224. https://doi.org/10.1016/bs.apcsb.2015.06.005.
10. Sumner, S., P. Söderhjelm, and U. Ryde. 2013. Effect of geometry optimizations on QM-Cluster and QM/MM studies of reaction energies in proteins. *J. Chem. Theory Comput.* 9:4205–4214.
11. Hu, L., P. Söderhjelm, and U. Ryde. 2013. Accurate reaction energies in proteins obtained by combining QM/MM and large QM calculations. *J. Chem. Theory Comput.* 9:640–649.
12. Hu, L., P. Söderhjelm, and U. Ryde. 2011. On the convergence of QM/MM energies. *J. Chem. Theory Comput.* 7:761–777.
13. Sumowski, C. V., and C. Ochsenfeld. 2009. A convergence study of QM/MM isomerization energies with the selected size of the QM region for peptidic systems. *J. Phys. Chem. A.* 113:11734–11741.
14. Liao, R. Z., and W. Thiel. 2013. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: a case study for acetylene hydratase. *J. Comput. Chem.* 34:2389–2397.
15. Solt, I., P. Kulhánek, …, M. Fuxreiter. 2009. Evaluating boundary dependent errors in QM/MM simulations. *J. Phys. Chem. B.* 113:5728–5735.
16. Vanpoucke, D. E. P., J. Oláh, …, G. Roos. 2015. Convergence of atomic charges with the size of the enzymatic environment. *J. Chem. Inf. Model.* 55:564–571.
17. Morgenstern, A., M. Jaszai, …, A. N. Alexandrova. 2017. Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chem. Sci. (Camb.).* 8:5010–5018.
18. Kulik, H. J. 2018. Large-scale QM/MM free energy simulations of enzyme catalysis reveal the influence of charge transfer. *Phys. Chem. Chem. Phys.* 20:20650–20660.
19. Alavi, F. S., M. Gheidi, …, U. Ryde. 2018. A novel mechanism of heme degradation to biliverdin studied by QM/MM and QM calculations. *Dalton Trans.* 47:8283–8291.
20. Hu, L., J. Eliasson, …, U. Ryde. 2009. Do quantum mechanical energies calculated for small models of protein-active sites converge? *J. Phys. Chem. A.* 113:11793–11800.
21. Rod, T. H., and U. Ryde. 2005. Quantum mechanical free energy barrier for an enzymatic reaction. *Phys. Rev. Lett.* 94:138302.
22. Rod, T. H., and U. Ryde. 2005. Accurate QM/MM free energy calculations of enzyme reactions: methylation by catechol O-methyltransferase. *J. Chem. Theory Comput.* 1:1240–1251.
23. Sharir-Ivry, A., R. Varatharaj, and A. Shurki. 2015. Challenges within the linear response approximation when studying enzyme catalysis and effects of mutations. *J. Chem. Theory Comput.* 11:293–302.
24. Karelina, M., and H. J. Kulik. 2017. Systematic quantum mechanical region determination in QM/MM simulation. *J. Chem. Theory Comput.* 13:563–576.
25. Di Paola, L., M. De Ruvo, …, A. Giuliani. 2013. Protein contact networks: an emerging paradigm in chemistry. *Chem. Rev.* 113:1598–1613.
26. Doncheva, N. T., K. Klein, …, M. Albrecht. 2011. Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.* 36:179–182.
27. Harris, T. V., and R. K. Szilagyi. 2016. Protein environmental effects on iron-sulfur clusters: a set of rules for constructing computational models for inner and outer coordination spheres. *J. Comput. Chem.* 37:1681–1696.
28. Zheng, M., and M. P. Waller. 2018. Yoink: an interaction-based partitioning API. *J. Comput. Chem.* 39:799–806.
29. National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. The National Academic Press, Washington, DC. https://doi.org/10.17226/25303.
30. Kanaan, N., J. J. Ruiz Pernía, and I. H. Williams. 2008. QM/MM simulations for methyl transfer in solution and catalysed by COMT: ensemble-averaging of kinetic isotope effects. *Chem. Commun. (Camb.).* 6114–6116:6114–6116.
31. Rod, T. H., P. Rydberg, and U. Ryde. 2006. Implicit versus explicit solvent in free energy calculations of enzyme catalysis: methyl transfer catalyzed by catechol O-methyltransferase. *J. Chem. Phys.* 124:174503.
32. Roca, M., V. Moliner, …, I. Tuñón. 2006. Activation free energy of catechol O-methyltransferase. Corrections to the potential of mean force. *J. Phys. Chem. A.* 110:503–509.
33. Hatstat, A. K., M. Morris, …, M. Cafiero. 2016. Ab initio study of electronic interaction energies and desolvation energies for dopaminergic ligands in the catechol-O-methyltransferase active site. *Comput. Theor. Chem.* 1078:146–162.
34. Yang, Z., R. Mehmood, …, H. J. Kulik. 2019. Revealing quantum mechanical effects in enzyme catalysis with large-scale electronic structure simulation. *React. Chem. Eng.* 4:298–315.
35. Roca, M., S. Martí, …, I. H. Williams. 2003. Theoretical modeling of enzyme catalytic power: analysis of "cratic" and electrostatic factors in catechol O-methyltransferase. *J. Am. Chem. Soc.* 125:7726–7737.
36. Roca, M., J. Andrés, …, J. Bertrán. 2005. On the nature of the transition state in catechol O-methyltransferase. A complementary study based on molecular dynamics and potential energy surface explorations. *J. Am. Chem. Soc.* 127:10648–10655.
37. García-Meseguer, R., K. Zinovjev, …, I. Tuñón. 2015. Linking electrostatic effects and protein motions in enzymatic catalysis. A theoretical analysis of catechol o-methyltransferase. *J. Phys. Chem. B.* 119:873–882.
38. Chen, X., and S. D. Schwartz. 2019. Examining the origin of catalytic power of catechol O-methyltransferase. *ACS Catal.* 9:9870–9879.
39. Patra, N., E. I. Ioannidis, and H. J. Kulik. 2016. Computational investigation of the interplay of substrate positioning and reactivity in catechol O-methyltransferase. *PLoS One.* 11:e0161868.
40. Lameira, J., R. P. Bora, …, A. Warshel. 2015. Methyltransferases do not work by compression, cratic, or desolvation effects, but by electrostatic preorganization. *Proteins.* 83:318–330.
41. Roca, M., V. Moliner, …, J. T. Hynes. 2006. Coupling between protein and reaction dynamics in enzymatic processes: application of Grote-Hynes Theory to catechol O-methyltransferase. *J. Am. Chem. Soc.* 128:6186–6193.
42. Saez, D. A., K. Zinovjev, …, E. Vöhringer-Martinez. 2018. Catalytic reaction mechanism in native and mutant catechol- O-methyltransferase from the adaptive string method and mean reaction force analysis. *J. Phys. Chem. B.* 122:8861–8871.
43. Jindal, G., and A. Warshel. 2016. Exploring the dependence of QM/MM calculations of enzyme catalysis on the size of the QM region. *J. Phys. Chem. B.* 120:9913–9921.
44. Ruggiero, G. D., I. H. Williams, …, I. Tuñón. 2004. QM/MM determination of kinetic isotope effects for COMT-catalyzed methyl transfer does not support compression hypothesis. *J. Am. Chem. Soc.* 126:8634–8635.
45. Kuhn, B., and P. A. Kollman. 2000. QM-FE and molecular dynamics calculations on catechol O- methyltransferase: free energy of activation in the enzyme and in aqueous solution and regioselectivity of the enzyme-catalyzed reaction. *J. Am. Chem. Soc.* 122:2586–2596.
46. Zhang, J., and J. P. Klinman. 2011. Enzymatic methyl transfer: role of an active site residue in generating active site compaction that correlates with catalytic efficiency. *J. Am. Chem. Soc.* 133:17134–17137.

47. Lautala, P., I. Ulmanen, and J. Taskinen. 2001. Molecular mechanisms controlling the rate and specificity of catechol O-methylation by human soluble catechol O-methyltransferase. *Mol. Pharmacol.* 59:393–402.
48. Yan, W., J. Zhou, …, B. Shen. 2014. The construction of an amino acid network for understanding protein structure and function. *Amino Acids.* 46:1419–1439.
49. Rutherford, K., I. Le Trong, …, W. W. Parson. 2008. Crystal structures of human 108V and 108M catechol O-methyltransferase. *J. Mol. Biol.* 380:120–130.
50. Word, J. M., S. C. Lovell, …, D. C. Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285:1735–1747.
51. Word, J. M., S. C. Lovell, …, D. C. Richardson. 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* 285:1711–1733.
52. Schrödinger, LLC. 2019. The PyMol molecular graphics system, version 2.3. Schrödinger, LLC, New York, NY.
53. Spicher, S., and S. Grimme. 2021. Single-point Hessian calculations for improved vibrational frequencies and rigid-rotor-harmonic-oscillator thermodynamics. *J. Chem. Theory Comput.* 17:1701–1714.
54. Dasgupta, S., and J. M. Herbert. 2020. Using atomic confining potentials for geometry optimization and vibrational frequency calculations in quantum-chemical models of enzyme active sites. *J. Phys. Chem. B.* 124:1137–1147.
55. Frisch, M. J., G. W. Trucks, …, D. J. Fox. 2016. Gaussian16, revision B.01. Gaussian Inc., Wallingford, CT.
56. Becke, A. D. 1993. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98:5648–5652.
57. Lee, C., W. Yang, and R. G. Parr. 1988. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B Condens. Matter.* 37:785–789.
58. Petersson, G. A., and M. A. Al-Laham. 1991. A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *J. Chem. Phys.* 94:6081–6090.
59. Hehre, W. J., R. Ditchfield, and J. A. Pople. 1972. Self—consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* 56:2257–2261.
60. Wadt, W. R., and P. J. Hay. 1985. Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *J. Chem. Phys.* 82:284–298.
61. Grimme, S., S. Ehrlich, and L. Goerigk. 2011. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* 32:1456–1465.
62. Barone, V., and M. Cossi. 1998. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A.* 102:1995–2001.
63. Cossi, M., N. Rega, …, V. Barone. 2003. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* 24:669–681.
64. Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 28:100–108.
65. RStudio Team. 2020. RStudio: integrated development environment for R. R Studio, PBC, Boston, MA.
66. Kassambara, A., and F. Mundt. 2020. factoextra: extract and visualize the results of multivariate data analyses http://www.sthda.com/english/rpkgs/factoextra.
67. Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol.* 63:411–423.
68. Jubb, H. C., A. P. Higueruelo, …, T. L. Blundell. 2017. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* 429:365–371.
69. Siegbahn, P. E. M., and M. R. A. Blomberg. 2000. Transition-metal systems in biochemistry studied by high-accuracy quantum chemical methods. *Chem. Rev.* 100:421–438.
70. Siegbahn, P. E. M., and T. Borowski. 2006. Modeling enzymatic reactions involving transition metals. *Acc. Chem. Res.* 39:729–738.