Towards a cheminformatic design for quantum mechanical enzyme models: the case of catechol-O-methyltransferase

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

*ndyonker@memphis.edu, (901)-678-2029

[a]Department of Chemistry, The University of Memphis, 213 Smith Chemistry Building, Memphis, Tennessee, 38152-3550, United States

[b]Department of Computer Science, The University of Memphis, 375 Dunn Hall Building, Memphis, Tennessee, 38152-3240, United States

[c]Department of Chemistry, Mississippi State University, 1120 Hand Lab, Mississippi State, Mississippi, 39762-9573, United States

T.J.S. ORCID: 0000-0002-4243-6078

Q.C. ORCID: 0000-0002-4640-2238

D.P. ORCID: 0000-0001-9521-4789

D.K.K. ORCID: 0000-0001-8043-3225

C.E.W. ORCID: 0000-0002-6917-2957

N.J.D. ORCID: 0000-0003-0435-2006

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; and T.J.S., Q.C., C.E.W., and N.J.D. wrote the paper.

Abstract

In order 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 (COMT), *RINRUS* was used to construct a series of QM-cluster models. The reactant, product, and transition state of the methyl transfer reaction was computed for a total of 527 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.

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 ONIOM modeling are complementary approaches that have leveraged advancements in quantum mechanical theory and molecular dynamics (MD) to continually increase the ubiquity of computational enzymology (4–6). As with all forms of modeling, the comparative accuracy of a model to reality is limited by the design of the model and relevant/reliable experimental data. For simulating the active site of enzymes, it is crucial to ensure 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). While 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.

While *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. While suitable models could be constructed this way, calibration studies have confirmed large spheres (and consequently large models) are needed for convergence of simulated enzyme thermodynamics/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, 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/Charge Shift Analysis (14, 24). However, such methods are essentially performing complicated computations on enzyme models in order to decide on an optimal model. Iterating residue selection processes to self-consistency 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 SI 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 API (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, minimizing trial and error using poorly or incorrectly designed models, and improving reproducibility of workflows and published results.

Additionally, enzyme models used to obtain insightful results must be reported in a reproducible manner, the simplest way being the inclusion of Cartesian coordinates for optimized structures in Supporting Information documents. The need for improved reporting

within the science community 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 highlight that reproducibility in QM/MM and QM-cluster modeling continues to be a problem, we conducted a survey of 58 computational QM/MM or QM-cluster model papers published within Jan 1 – Mar 31 of 2015 and Jan 1 – Mar 31 2019 to evaluate whether the models could be directly reproduced via reporting of Cartesian coordinates (see SI for details). Our survey indicated only 20 papers (34%) reported Cartesian coordinates to the extent that reproduction is possible. Given the absence of consistent community reporting, embedding reproducibility into a systematic model design workflow would be a large step towards research standards in computational enzymology. Future transformative leaps in computational biochemical method development will be severely hindered without ontologies, rigorous calibration efforts, and shared best practices. We argue that *RINRUS* is the first community-oriented software to facilitate rational, reproducible, and rigorous QM-cluster model workflows and models.

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/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 *S*-adenosylmethionine (SAM) coenzyme to the oxygen of a Mg^{2+}-bound catecholate substrate (CAT, Figure 1A). 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/kinetics requires radial QM-regions of 400 - 600 atoms (8, 18, 34). Unfortunately, conventional 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/thermodynamic properties should converge quickly as the size of the QM-region grows in a QM/MM partition. Slow convergence behavior of COMT has been attributed to the non-spherical active site, requiring an accurate description of both the Mg^{2+}/catechol coordination chemistry and the electrostatic stabilization of the large SAM cofactor (34).

While 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. Using the workflow employed by *RINRUS*, we present the reaction thermodynamics and free energies of activation for hundreds of QM-cluster models of COMT. The goal is to identify inter-residue contact features that predictively construct accurate and efficient QM-cluster models of enzymes

different than that of COMT. Our cheminformatics perspective will be the first rigorous step towards a translatable and generalized computational enzymology protocol.

Results and Discussion

We began by computing a contact-based residue interaction network (Figure 1B) for an X-ray crystal structure of human COMT (Protein Data Bank ID 3BWM), which indicated 27 protein residues and 4 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 SI). 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, 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 shell" maximal model that includes all residues with quantified contacts with any of the seed fragments. This maximal model is ellipsoidal in shape (Figure 4B), reflective of the non-spherical 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 SI. In total, the methyl transfer transition state and connecting reactants/products for 527 unique QM-cluster models were computed. 1581 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 contact interactions with the seed but 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 (Figure 2A). 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 largest distance-based models (containing 444 and 447 atoms) incorrectly predict an endergonic reaction. Qualitatively incorrect thermodynamics corresponds to the addition of the charged residue K46, which as previously noted, does not have direct contact interactions with the seed. At best, the addition of peripheral, non-interacting residues adds unnecessary time to the DFT simulations, as observed with the addition of the uncharged

N92 residue changing $\Delta G^{\ddagger}$ and ΔG_{rxn} by < 0.2 kcal/mol in the 486-atom model. However, the charged K46 side chain distorts the simulated microenvironment in the absence of counterbalancing residue(s). Without a proper method for adding residues in a physically meaningful way, the distance-based scheme has no way to adapt to this predicament beyond further undirected expansion of the largest models.

As a step towards 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: 11.7 kcal/mol for $\Delta G^{\ddagger}$ and −5.9 kcal/mol for ΔG_{rxn}. The converged reference value for $\Delta G^{\ddagger}$ 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. With an improved ranking scheme using number of residue-seed contacts, $\Delta G^{\ddagger}$ and ΔG_{rxn} both converge by the 302-atom model (Figure 2B). While 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 where 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 SI 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 QM-cluster model size is greater than ~300 atoms (Figure 2C), 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 towards 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/substrate protonation states, orientation of explicit solvent molecules, and conformational sampling (7, 9). While 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 (Figure 2C) were reconstructed by-hand by the PI. 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. The results of these “bespoke” models are presented in Figure 2D and are shown to be comparable to the models built by *RINRUS* (Figure 2C). 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 conformers, 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 which is founded on a traceable cheminformatic basis and a reproducible, rational workflow.

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 either residue main chain, residue side chain, or explicit water molecule) and the contact type (wide contact, close contact, small overlap, big overlap, 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 SI for details). To further increase the number of models and dataset size, the sets of residues classified

by contact types were repartitioned into a second combinatoric approach (Combinatoric Scheme 2, see SI for details), though the generation of these sets is not rigorous or necessarily applicable to other biosystems. Given the limitations of time and resources, 91 (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 (Figure S5). As the goal is identifying small, yet accurate, QM-cluster models, the cost of expanding the dataset 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 (Figure 3A and B), a 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, since the root mean square deviation (RMSD) 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 (Figure S4; standard deviation, 0.18 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 Figure 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 models, the *k*-means clustering algorithm was used to partition the entire dataset of unique QM-cluster models into six groups (Figure 3C) based upon their

predicted $\Delta G^{\ddagger}$ and ΔG_{rxn} (48). 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 (Figures 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 E199 and only 11% contain M40. Without these residues, the QM-cluster models are missing 1) the stabilizing hydrogen bonding interactions between E199 and the catechol and 2) the hydrophobic interactions between M40 and the SAM, resulting in consistently poor accuracy 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 the catechol, but plays a role in stabilizing and properly orienting the SAM. Other residues apart from the eight illustrated in Figure 4 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 if 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 (Figure 2B) 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, 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 Figure 2B where 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 to the interatomic contacts (*e.g.*, through *Arpeggio* (49)), to developing a weighting system to favor certain contact interactions (*e.g.*, hydrogen bonding, polar, 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. While computational resources and scaling limitations of quantum chemistry are among factors limiting progress in this field, little attention has been given towards 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 non-spherical 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 towards 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/classifying them. The catalytic methyl transfer reaction of the human COMT enzyme was simulated with a total of 527 unique models, illustrating how information from *RINRUS* were 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. Ranking residues by the frequency of their contacts with the seed demonstrated particular usefulness, 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 Density Functional Theory, 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 are 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 expanding into automated 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/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 towards discovering and implementing a computationally inexpensive, cheminformatic-based means for constructing reproducible, rational, and rigorous enzyme models. Admittedly, this single case study does not fully address all parameters of enzyme QM-cluster model construction and centers around one out of countless possible enzymes. Nevertheless,

reproducible workflows in computational enzymology, supported by *RINRUS* development, will improve openness, data sharing, and facilitate novel cyber- and software infrastructure in biochemistry and biology.

Methods

The initial atomic coordinates for building the models were taken from an X-ray crystal structure of COMT (PDB ID: 3BWM) containing the coenzyme *S*-adenosyl methionine (SAM) and a 3,5-dinitrocatechol inhibitor coordinated to the active site metal (50). Hydrogens were added to this protein structure using the program *Reduce* (51), and the two nitro-groups of 3,5-dinitrocatechol were replaced with hydrogens to form the catechol (CAT) substrate. The program *Probe* (52) was used to roll a small spherical probe over the van der Waals surface of this modified structure to identify and classify non-covalent interatomic contact interactions. This information was compiled into an interaction network (see SI) for identifying inter-residue contact interactions. Focusing on the chemically reactive species for COMT (SAM, CAT, and the Mg^{2+} CAT binds to), a total of 27 amino acids and 4 crystallographic waters are predicted to have interatomic contact interactions with this seed.

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, HOH411; Figure 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 have 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 (53). To maintain the general shape and semi-rigid 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. Further details about residue selection and model trimming are provided in the SI. Though other research groups who employ QM-cluster models may have developed internal research protocols for trimming residues/fragments and freezing backbone atoms, we intend *RINRUS* to be the first 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 (54). The models were geometrically optimized using density functional theory (DFT) with the hybrid B3LYP exchange-correlation functional (55, 56). The computations used the 6-31G(d') basis set for N, O, and S (57); the 6-31G basis set for C and H atoms (58); and the LANL2DZ effective core potential and basis set combination for Mg (59). The Grimme D3 (Becke-Johnson) dispersion correction (GD3BJ) was also included (60) along with a conductor-like polarizable continuum model (CPCM) using UAKS sets of atomic radii, a nondefault electronic scaling factor of 1.2, and a dielectric constant of $\varepsilon = 4$ (61, 62). Unscaled harmonic vibrational frequency calculations were used to confirm all stationary points as either minima or transition states. Stationary points were found by first pre-optimizing the model to the reactant structure. This pre-optimized 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 (48) was run through *RStudio v.3.6.3* (63). Elbow and gap statistics (Figure S6) were used to identify a $k = 6$ for the cluster analysis (64).

Acknowledgements

The authors thank Prof. 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 No. 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.). This support does not necessarily imply endorsement by the University of research conclusions. The High Performance Computing Center and the Computational Research on Materials Institute at The University of Memphis (CROMIUM) also provided generous resources for this research.

References

1. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, K. N. Houk, Computational enzyme design. *Angew. Chemie - Int. Ed.* **52**, 5700–5725 (2013).
2. P. A. Kollman, B. Kuhn, M. Peräkylä, 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 (2002).
3. The Nobel Prize in Chemistry 2013. *R. Swedish Acad. Sci.* (2013) (June 9, 2020).
4. S. Ahmadi, *et al.*, Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *Int. J. Quantum Chem.* **118**, e25558 (2018).

5. K. Kmita, *et al.*, Accessory NUMM (NDUFS6) subunit harbors a Zn-binding site and is essential for biogenesis of mitochondrial complex I. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5685–5690 (2015).

6. X. Li, P. E. M. Siegbahn, U. Ryde, Simulation of the isotropic EXAFS spectra for the S2 and S3 structures of the oxygen evolving complex in photosystem II. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3979–3984 (2015).

7. R. Lonsdale, J. N. Harvey, A. J. Mulholland, A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **41**, 3025–3038 (2012).

8. H. J. Kulik, J. Zhang, J. P. Klinman, T. J. Martínez, How large should the QM region be in QM/MM calculations? the case of catechol O-methyltransferase. *J. Phys. Chem. B* **120**, 11381–11394 (2016).

9. T. Borowski, M. Quesne, M. Szaleniec, QM and QM/MM methods compared: Case studies on reaction mechanisms of metalloenzymes in *Advances in Protein Chemistry and Structural Biology*, (Academic Press Inc., 2015), pp. 187–224.

10. S. Sumner, P. Söderhjelm, U. Ryde, Effect of geometry optimizations on QM-Cluster and QM/MM studies of reaction energies in proteins. *J. Chem. Theory Comput.* **9**, 4205–4214 (2013).

11. L. Hu, P. Söderhjelm, U. Ryde, Accurate reaction energies in proteins obtained by combining QM/MM and large QM calculations. *J. Chem. Theory Comput.* **9**, 640–649 (2013).

12. L. Hu, P. Söderhjelm, U. Ryde, On the convergence of QM/MM energies. *J. Chem. Theory Comput.* **7**, 761–777 (2011).

13. C. V. Sumowski, C. Ochsenfeld, 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 (2009).

14. R. Z. Liao, W. Thiel, Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *J. Comput. Chem.* **34**, 2389–2397 (2013).

15. I. Solt, *et al.*, Evaluating boundary dependent errors in QM/MM simulations. *J. Phys. Chem. B* **113**, 5728–5735 (2009).

16. D. E. P. Vanpoucke, J. Oláh, F. De Proft, V. Van Speybroeck, G. Roos, Convergence of atomic charges with the size of the enzymatic environment. *J. Chem. Inf. Model.* **55**, 564–571 (2015).

17. A. Morgenstern, M. Jaszai, M. E. Eberhart, A. N. Alexandrova, Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chem. Sci.* **8**, 5010–5018 (2017).

18. H. J. Kulik, Large-scale QM/MM free energy simulations of enzyme catalysis reveal the influence of charge transfer. *Phys. Chem. Chem. Phys.* **20**, 20650–20660 (2018).

19. F. S. Alavi, M. Gheidi, M. Zahedi, N. Safari, U. Ryde, A novel mechanism of heme

degradation to biliverdin studied by QM/MM and QM calculations. *Dalt. Trans.* **47**, 8283–8291 (2018).

20. L. Hu, J. Eliasson, J. Heimdal, U. Ryde, Do quantum mechanical energies calculated for small models of protein-active sites converge. *J. Phys. Chem. A* **113**, 11793–11800 (2009).

21. T. H. Rod, U. Ryde, Quantum mechanical free energy barrier for an enzymatic reaction. *Phys. Rev. Lett.* **94**, 1–4 (2005).

22. T. H. Rod, U. Ryde, Accurate QM/MM free energy calculations of enzyme reactions: Methylation by catechol O-methyltransferase. *J. Chem. Theory Comput.* **1**, 1240–1251 (2005).

23. A. Sharir-Ivry, R. Varatharaj, A. Shurki, Challenges within the linear response approximation when studying enzyme catalysis and effects of mutations. *J. Chem. Theory Comput.* **11**, 293–302 (2015).

24. M. Karelina, H. J. Kulik, Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **13**, 563–576 (2017).

25. L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, A. Giuliani, Protein contact networks: An emerging paradigm in chemistry. *Chem. Rev.* **113**, 1598–1613 (2013).

26. N. T. Doncheva, K. Klein, F. S. Domingues, M. Albrecht, Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.* **36**, 179–182 (2011).

27. T. V. Harris, R. K. Szilagyi, 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 (2016).

28. M. Zheng, M. P. Waller, Yoink: An interaction-based partitioning API. *J. Comput. Chem.* **39**, 799–806 (2018).

29. National Academies of Sciences, Engineering, and Medicine, Reproducibility and Replicability in Science. Washington, DC: The National Academic Press (2019).

30. N. Kanaan, J. J. Ruiz Pernía, I. H. Williams, QM/MM simulations for methyl transfer in solution and catalysed by COMT: Ensemble-averaging of kinetic isotope effects. *Chem. Commun.*, 6114–6116 (2008).

31. T. H. Rod, P. Rydberg, U. Ryde, Implicit versus explicit solvent in free energy calculations of enzyme catalysis: Methyl transfer catalyzed by catechol O -methyltransferase. *J. Chem. Phys.* **124**, 174503 (2006).

32. M. Roca, V. Moliner, J. J. Ruiz-Pernía, E. Silla, I. Tuñón, Activation free energy of catechol O-methyltransferase. Corrections to the potential of mean force. *J. Phys. Chem. A* **110**, 503–509 (2006).

33. A. K. Hatstat, M. Morris, L. W. Peterson, M. Cafiero, 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 (2016).

34. Z. Yang, *et al.*, Revealing quantum mechanical effects in enzyme catalysis with large-scale electronic structure simulation. *React. Chem. Eng.* **4**, 298–315 (2019).

35. M. Roca, *et al.*, Theoretical modeling of enzyme catalytic power: Analysis of "cratic" and electrostatic factors in catechol O-methyltransferase. *J. Am. Chem. Soc.* **125**, 7726–7737 (2003).

36. M. Roca, J. Andrés, V. Moliner, I. Tuñón, J. Bertrán, 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 (2005).

37. R. García-Meseguer, K. Zinovjev, M. Roca, J. J. Ruiz-Pernía, I. Tuñón, Linking electrostatic effects and protein motions in enzymatic catalysis. A theoretical analysis of catechol O-methyltransferase. *J. Phys. Chem. B* **119**, 873–882 (2015).

38. X. Chen, S. D. Schwartz, Examining the Origin of Catalytic Power of Catechol O-Methyltransferase. *ACS Catal.* **9**, 9870–9879 (2019).

39. N. Patra, E. I. Ioannidis, H. J. Kulik, Computational Investigation of the Interplay of Substrate Positioning and Reactivity in Catechol O-Methyltransferase. *PLoS One* **11**, e0161868 (2016).

40. J. Lameira, R. P. Bora, Z. T. Chu, A. Warshel, Methyltransferases do not work by compression, cratic, or desolvation effects, but by electrostatic preorganization. *Proteins Struct. Funct. Bioinforma.* **83**, 318–330 (2015).

41. M. Roca, V. Moliner, I. Tuñón, J. T. Hynes, 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 (2006).

42. D. A. Saez, K. Zinovjev, I. Tuñón, E. Vöhringer-Martinez, 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 (2018).

43. G. Jindal, A. Warshel, Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **120**, 9913–9921 (2016).

44. G. D. Ruggiero, I. H. Williams, M. Roca, V. Moliner, I. Tuñón, QM/MM determination of kinetic isotope effects for COMT-catalyzed methyl transfer does not support compression hypothesis. *J. Am. Chem. Soc.* **126**, 8634–8635 (2004).

45. B. Kuhn, P. A. Kollman, 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 (2000).

46. J. Zhang, J. P. Klinman, 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 (2011).

47. P. Lautala, I. Ulmanen, J. Taskinen, Molecular Mechanisms Controlling the Rate and Specificity of Catechol O-Methylation by Human Soluble Catechol O-Methyltransferase.

Mol. Pharmacol. **59**, 393-402 (2001).

48. J. A. Hartigan, M. A. Wong, Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28**, 100 (1979).

49. H. C. Jubb, *et al.*, Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **429**, 365–371 (2017).

50. K. Rutherford, I. Le Trong, R. E. Stenkamp, W. W. Parson, Crystal Structures of Human 108V and 108M Catechol O-Methyltransferase. *J. Mol. Biol.* **380**, 120–130 (2008).

51. J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747 (1999).

52. J. M. Word, *et al.*, Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–33 (1999).

53. The PyMOL Molecular Graphics System, Version 2.3. Schrödinger, LLC.

54. M. J. Frisch, *et al.*, Gaussian16 (Revision B.01), Gaussian Inc. Wallingford CT. (2016).

55. A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648 (1993).

56. C. Lee, W. Yang, R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).

57. G. A. Petersson, M. A. Al-Laham, 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 (1991).

58. W. J. Hehre, R. Ditchfield, J. A. Pople, Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **56**, 2257–2261 (1972).

59. W. R. Wadt, P. J. Hay, Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *J. Chem. Phys.* **82**, 284–298 (1985).

60. S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–65 (2011).

61. V. Barone, M. Cossi, Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **102**, 1995–2001 (1998).

62. M. Cossi, N. Rega, G. Scalmani, V. Barone, Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **24**, 669–681 (2003).

63. RStudio Team, RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA (2019).

64. R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411–423 (2001).

Figure 1. (A) COMT catalyzes the methyl-transfer reaction from *S*-adenosylmethionine (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).

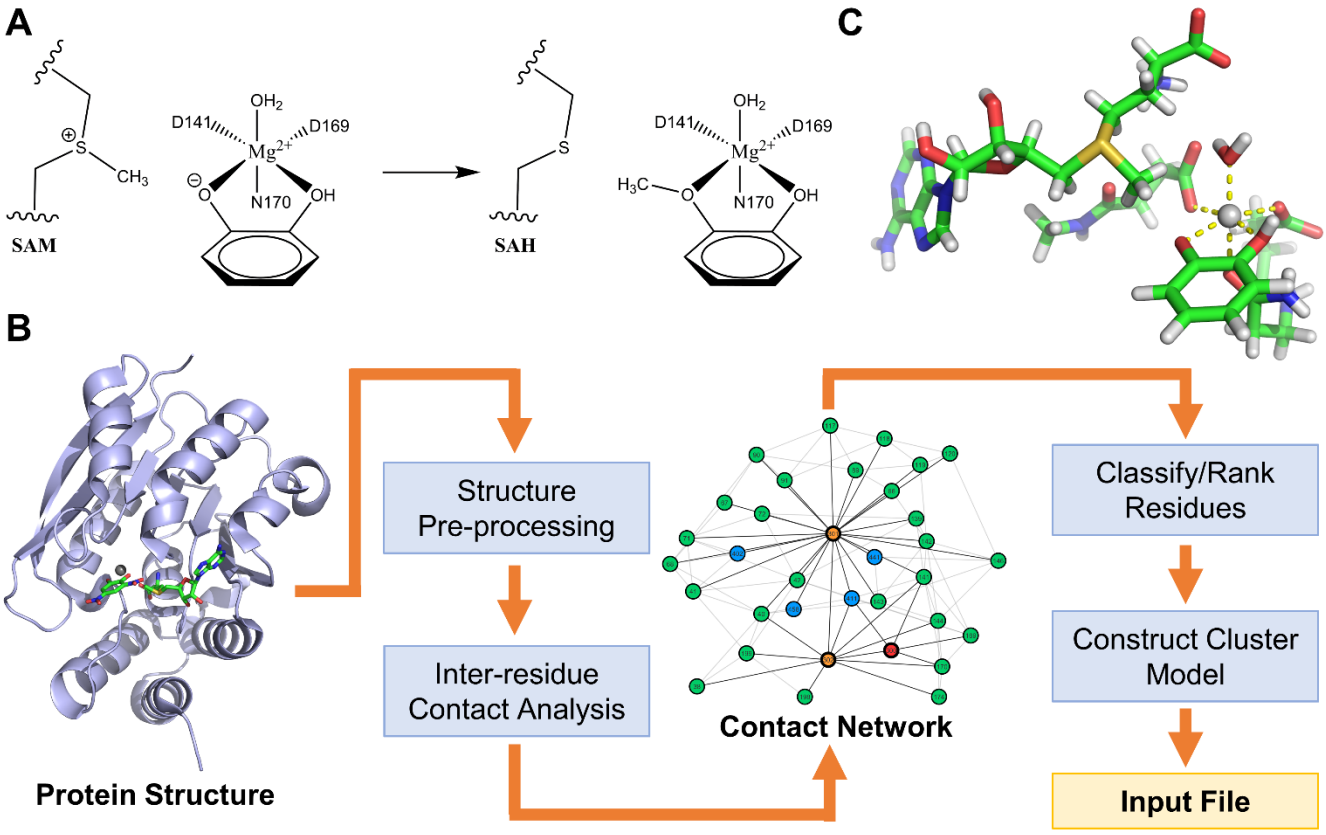

Figure 2. Computed methyl transfer $\Delta G^{\ddagger}$ (circle) and $\Delta G^{\ddagger}_{rxn}$ (triangle) free energies as models are systematically built-up through different methods of ranking residues including distance from the seed (A), total number of contacts with the seed (B), frequency of residue in Combinatoric Scheme 2 sets (C), and a by-hand reconstruction of models by frequency of residue in Combinatoric Scheme 2 sets (D). Grey lines indicate the reference convergence values.

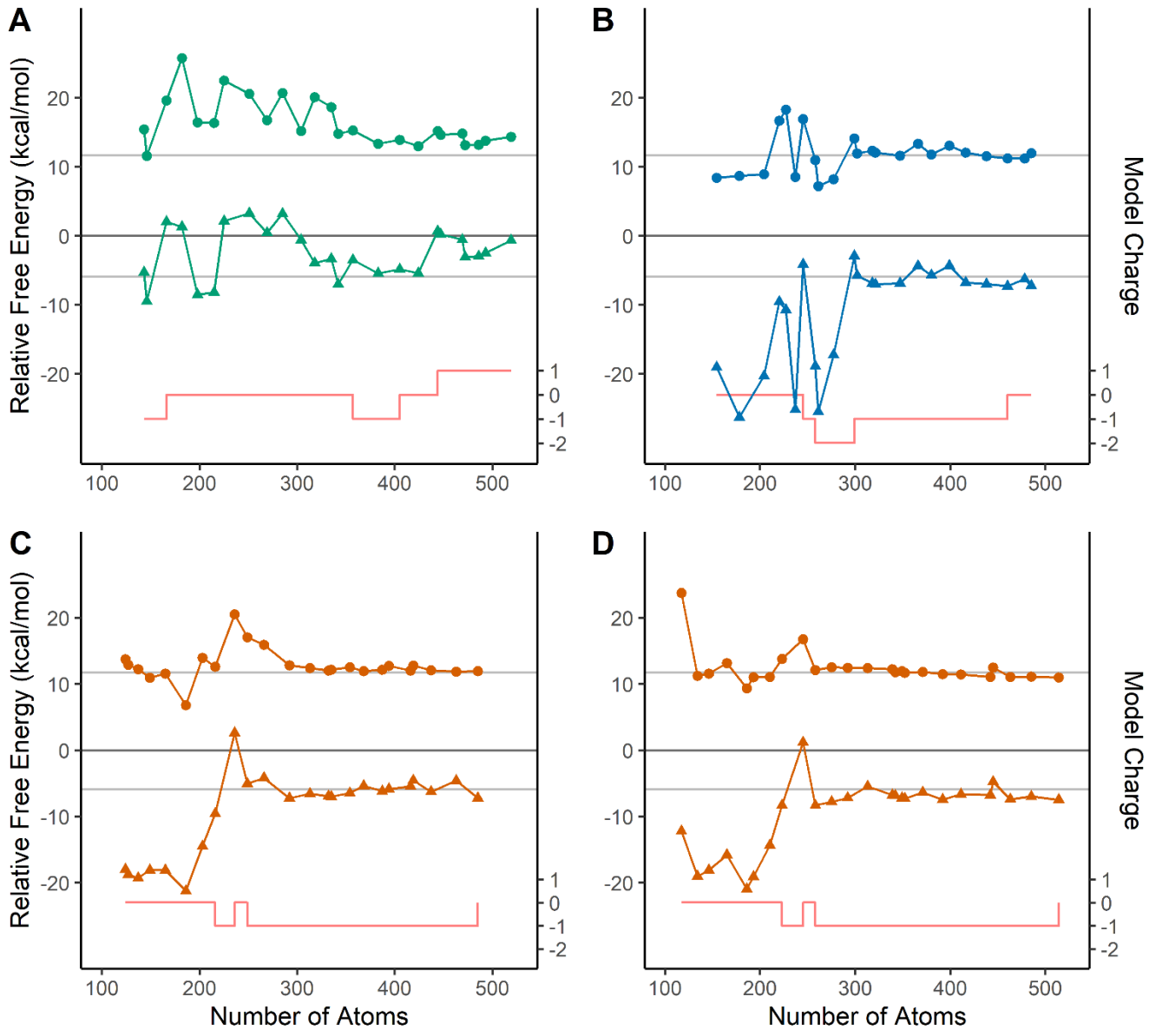
A
B
C
D
Relative Free Energy (kcal/mol)
Model Charge
Number of Atoms
20
10
0
-10
-20
1
0
-1
-2
100
200
300
400
500

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. Grey lines indicate the reference convergence values.

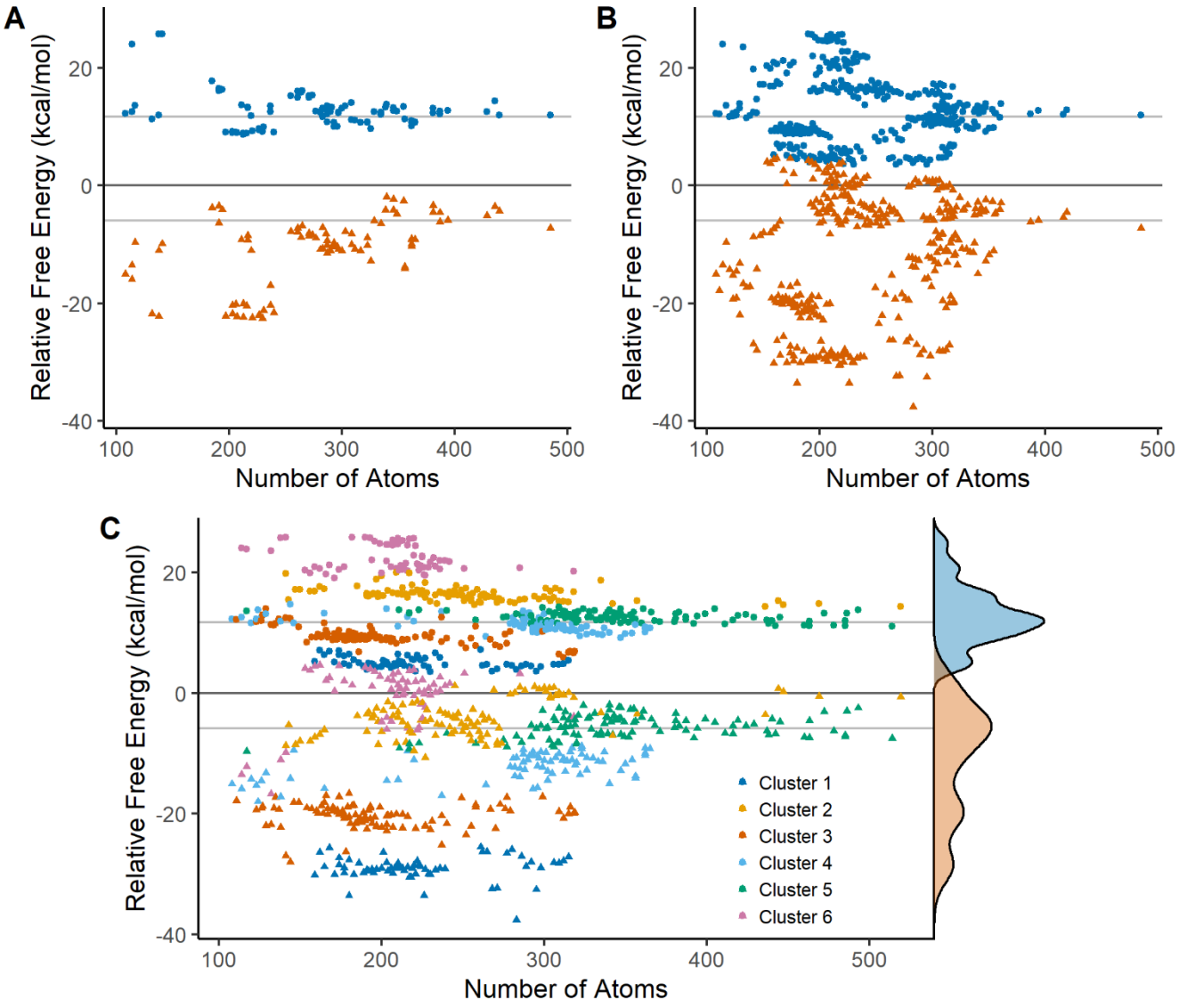

Figure 4. A) Relative frequency for each residue being present in the models of a *k*-cluster. Values are proportionally shaded to emphasize differences in residue composition among *k*-clusters. B) Visualization of the maximal 485-atom model highlighting the residues that occur in >80% of Cluster 5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