PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Correlating enzymatic reactivity for different substrates using transferable data-driven collective variables

Sudip Das[a], Umberto Raucci[a], Rui P. P. Neves[b], Maria J. Ramos[b,1], and Michele Parrinello[a,1]

Affiliations are included on p. 6.

Contributed by Michele Parrinello; received August 15, 2024; accepted October 27, 2024; reviewed by Michele Cascella and Pratyush Tiwary

Machine learning (ML) is transforming the investigation of complex biological processes. In enzymatic catalysis, one significant challenge is identifying the reactive conformations (RC) of the enzyme:substrate complex where the substrate assumes a precise arrangement in the active site necessary to initiate a reaction. Traditional methods are hindered by the complexity of the multidimensional free energy landscape involved in the transition from nonreactive to reactive conformations. Here, we applied ML techniques to address this challenge, focusing on human pancreatic α-amylase, a crucial enzyme in type-II diabetes treatment. Using ML-based collective variables (CVs), we correlated the probability of being in a RC with the experimental catalytic activity of several malto-oligosaccharide substrates. Our findings demonstrate a remarkable transferability of these CVs across various compounds, significantly streamlining the modeling process and reducing both computational demand and manual intervention in setting up simulations for new substrates. This approach not only advances our understanding of enzymatic processes but also holds substantial potential for accelerating drug discovery by enabling rapid and accurate evaluation of drug efficacy across different generations of inhibitors.

machine learning-based collective variables | transfer learning | active site and substrate pre-organization | enzyme catalysis | glycolysis

Due to its complexity, enzymatic reactivity is an area of research in which machine learning (ML) methods can make a difference despite the amazing diversity that both the enzymatic systems, their substrate counterparts, and the consequent reaction mechanisms display (1, 2). Indeed, ML-based techniques are powerful tools when addressing multifaceted processes that encompass a large number of coupled degrees of freedom. For instance, in the case of enzyme catalysis, the substrate requires a precise arrangement in the active site to initiate a reaction, and thus, only a fraction of conformations within the enzyme:substrate complex ensemble are catalytically active (3–15). Measuring this probability is hard, but this information is fundamental to explain the different catalytic activity of an enzyme toward different substrates.

When the probability of finding the reactants in a reactive conformation (RC) is high, the enzymatic reactivity should similarly be high, regardless of the substrate (Fig. 1*A*). This assumption is valid if the rate of the chemical transformation is consistent across different substrates of the same class. The ability to predict and understand when and how RC occur can fundamentally enhance our grasp of enzymatic reactivity (6, 16), and in this context, ML techniques can significantly accelerate the process.

Recently, using state-of-the-art enhanced simulation approaches coupled with ML techniques, we have explored the catalytic reaction space and characterized the transformation from nonreactive conformations (NRC) to RC of the human pancreatic α-amylase (HPA) (6), a major target for the drug treatment of type-II diabetes (18, 19).

Our endeavor required capturing the full complexity of the transformation from NRC to RC, which included multiple coupled degrees of freedom, such as intermolecular contacts and water coordination. This was achieved through the combination of a highly efficient enhanced sampling approach, like On-the-fly Probability Enhanced Sampling (OPES) (20, 21), with data-driven ML-based collective variables (ML-CVs) (22–33) trained to capture the slow modes of the system. Selecting the appropriate low-dimensional CVs is challenging for complex biological systems. Intuition-derived parameters, such as distances or water coordination, can initially distinguish between reactive and nonreactive states of the enzyme:substrate complex. However, accurately describing their transformation requires capturing the complex interplay among multiple coupled degrees of freedom. This can be achieved through ML-CVs that express the effects of a large set of descriptors (e.g., distances and dihedral angles) in a nonlinear manner.

Significance

While data-driven collective variables (CVs) have been applied to various complex biological problems, such as ligand binding, their transferability across different substrates has not been thoroughly investigated. Here, we show that machine learning-based CVs (ML-CVs), when trained with the appropriate set of descriptors, can be effectively transferred to simulate the dynamics of other systems that exhibit the same binding poses. This approach allowed bypassing the training process of specific CVs for each substrate, which typically requires extensive knowledge about the system and significant user involvement to select meaningful physical descriptors. This study represents a first step toward developing a generalized protocol for designing transferable ML-CVs capable of describing ligands with different scaffolds and binding modes.

Author contributions: S.D., U.R., M.J.R., and M.P. designed research; S.D. and U.R. performed research; S.D., U.R., and R.P.P.N. analyzed data; and S.D., U.R., R.P.P.N., M.J.R., and M.P. wrote the paper.

Reviewers: M.C., Universitetet i Oslo; and P.T., University of Maryland at College Park.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: mjramos@fc.up.pt or michele.parrinello@iit.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2416621121/-/DCSupplemental.

Published November 26, 2024.

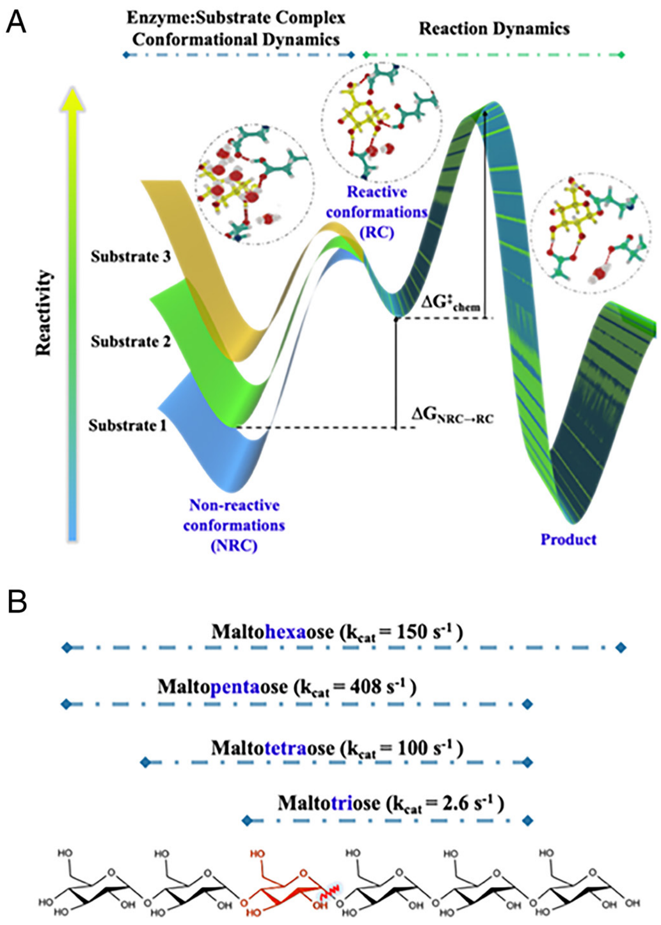
A
Enzyme:Substrate Complex Conformational Dynamics
Reaction Dynamics
Reactivity
Substrate 3
Substrate 2
Substrate 1
Reactive conformations (RC)
$\Delta G^{\ddagger}_{chem}$
$\Delta G_{NRC \rightarrow RC}$
Non-reactive conformations (NRC)
Product
B
Maltohexaose (k_{cat} = 150 s^{-1})
Maltopentaose (k_{cat} = 408 s^{-1})
Maltotetraose (k_{cat} = 100 s^{-1})
Maltotriose (k_{cat} = 2.6 s^{-1})
HO
OH

Fig. 1. (*A*) Schematic representation of the complexity of the free energy landscape in enzymatic catalysis, showing how the probability of the enzyme:substrate complex adopting a RC correlates with catalytic reactivity, for substrates showing the same activation energy for the chemical step. (*B*) Schematic of the sugar substrates alongside their experimental catalytic rate (17). The bond targeted for cleavage during catalysis is highlighted with a red curvy line. This bond was experimentally found to be the most probable cite of hydrolysis (16). The sugar unit highlighted in red is referred to as the reactive sugar ring in the text.

Using such ML-CVs we were able to obtain a detailed understanding of the transformation from NRC to RC of HPA, capturing the full complexity of the process (6). However, a challenge with ML approaches lies in their system specificity: Initially trained on a particular system, their direct applicability to others within the same class is not obvious.

Here, we show that with a careful selection of input descriptors for the neural network, our ML tools exhibit a remarkable transferability across various HPA substrates within the same class. This transferability allows us to significantly reduce both the computational time and manual intervention in setting up simulations for new substrates. This is essential for accelerating the understanding of the complex reactivity trends of HPA, and elucidating the reorganization of the Michaelis complex in various substrates.

Results and Discussion

HPA catalyzes the hydrolysis of α-(1, 4)-glycosidic linkages in a wide variety of starches, including various malto-oligosaccharides (34). Using our data-driven CVs, we analyzed four such oligosaccharides: maltotriose, maltotetraose, maltopentaose, and maltohexaose, which consist of three, four, five, and six glucose units, respectively (Fig. 1*B*). The experimentally determined k_{cat} exhibits a nonlinear relation with sugar size and shape (16) (Fig. 1*B*). We have correlated the variation of activity of HPA in the four experimentally investigated malto-oligosaccharides, with the respective probability to be in a RC. Indeed, given that the mechanisms for chemical conversion from reactant to product catalyzed by HPA are thought to be similar across these substrates (16), the observed differences in experimental catalytic rates may stem from their distinct binding abilities and mechanisms to reach catalytically active conformations of the Michaelis complex.

The reaction mechanism for the hydrolysis of the α-(1, 4)-glycosidic bond has been thoroughly documented in the literature (35–37). The rate-limiting step involves a nucleophilic attack by the Asp197 carboxylate group on the C_1 carbon of the sugar, leading to the cleavage of the glycosidic C_1–O_{gly} bond (Fig. 2 *A* and *B*). Concurrently, a proton transfer occurs between the acidic Glu233 and the glycosidic oxygen, O_{gly}. This proton transfer is facilitated by a buried water molecule (W_1), coordinating Glu233 (10). This mechanism leads to the formation of a covalently bound enzyme–substrate complex.

We studied the binding of a maltopentaose substrate with HPA (6) and identified three different substrate-binding modes, NRC_1, NRC_2, and RC (Fig. 2*C*) [referred to as states C, B, and A, respectively, in earlier study (6)]. In the binding mode NRC_1, the reactive sugar ring is positioned far from the nucleophile (approximately 9 Å) and it is solvated by ten water molecules, binding loosely to the active site in a NRC. Then, the substrate displaces some of the waters from the active site cavity and forms some of the reactive contacts with the enzyme catalytic residues, reaching the intermediate conformation NRC_2. This is also a NRC as it prevents catalysis via the formation of a hydrogen bond between the nucleophile Asp197 and the acidic Glu233. Finally, the substrate fully occupies the active site cleft and the reactive contacts between the enzyme and the substrate are maximized (6, 10, 35–37) reaching the substrate-binding mode RC, and only RC can follow the mechanism described above (Fig. 2*B*).

Subsequently, using deep targeted discriminant analysis (Deep-TDA) (38), we trained two CVs (Deep-Contact CV incorporating the enzyme–substrate reactive contacts and Deep-Water CV taking care of the water environment around the reactive sugar ring) to differentiate RC and NRC and to identify these three distinct binding modes of maltopentaose with HPA. Technical details about the construction of Deep-Contact and Deep-Water CVs are provided in the *Materials and Methods* section.

In the hope of bypassing the lengthy task of building CVs for the binding modes of HPA to maltotriose, maltotetraose, and maltohexaose, we first checked whether these Deep-Contact and Deep-Water CVs trained for the HPA-maltopentaose system were able to discriminate between all the systems and eventually be used to calculate their free energy surfaces (FES). To do so, we performed classical unbiased molecular dynamics simulations on the Michaelis complex of HPA with those substrates. Similar to what we found previously for maltopentaose, the most probable path leading to the reactive conformations of the complex was also composed of three distinct conformations (*SI Appendix*, Fig. S1). Only for the maltotriose substrate, we observed additional minima, which likely resulted from its smaller volume relative to the catalytic pocket and the accessibility of its reactive sugar ring to water (Fig. 1*B*). For the same reason, the water content in both the NRC_2 and RC states of the HPA–maltotriose complex is higher compared to that in complexes with other sugars (*SI Appendix*, Fig. S1).

To verify the consistency of these states across different systems, we also analyzed them using intuition-derived CVs, namely the distance from the center of mass of the reactive glucose ring to the C_α of the nucleophilic residue Asp197, and the coordination number

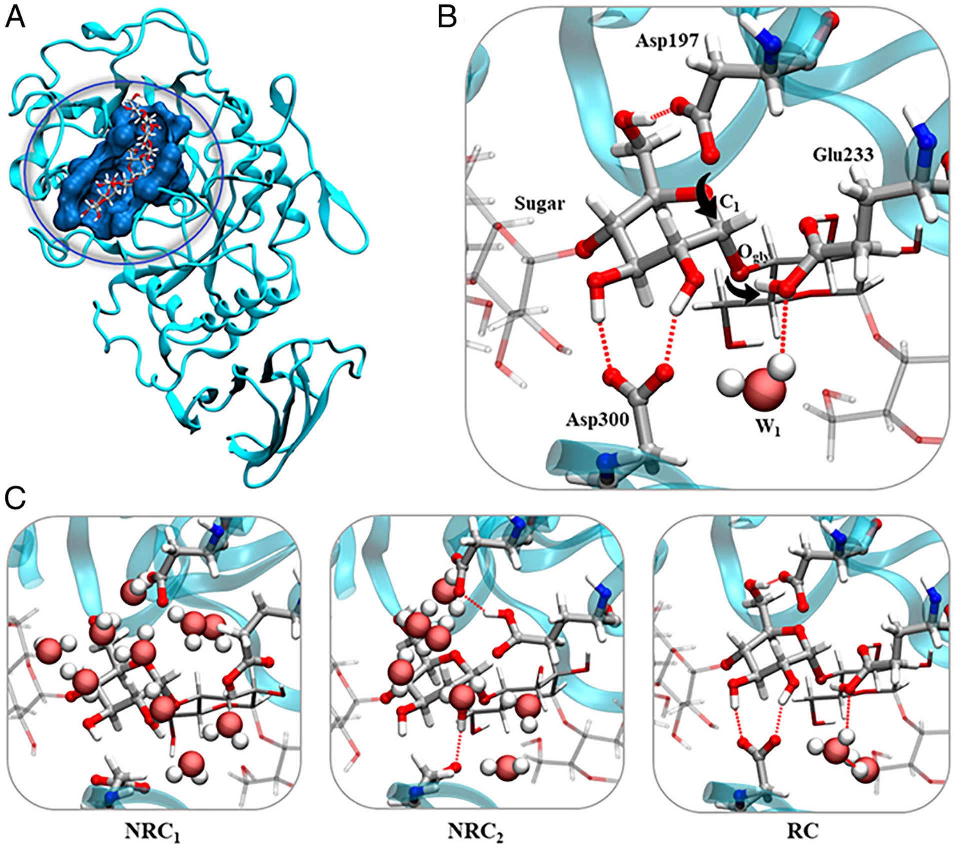

Fig. 2. (*A*) HPA bound to maltopentaose substrate. The active site cavity of HPA is shown as a blue surface. (*B*) Reaction mechanism for the rate-limiting step of sugar degradation catalyzed by HPA. (*C*) Different conformations of the HPA:maltopentaose Michaelis complex. The water molecules solvating the reactive sugar ring are also shown.

of water molecules to the reactive glucose ring (*SI Appendix*, Fig. S2). The chemical features were qualitatively maintained across the systems, confirming that our deep learned CVs are able to correctly capture the physics of the distinct binding poses and are transferable among the four substrates.

Finally, we used the OPES method with the maltopentaose-derived Deep-Contact and Deep-Water CVs, and, remarkably, we have been able to determine the FESs for the transformations from NRC to RC of each of the four malto-oligosaccharides complexed with HPA (Fig. 3).

From state NRC_1 to NRC_2 ($\Delta G_{\mathrm{NRC1 \rightarrow NRC2}}$), entropy is the major player, favoring the release of the active site water into the bulk solvent and, thus, there is a correlation between the size of the molecule and entropy gain, which favors the largest substrate,

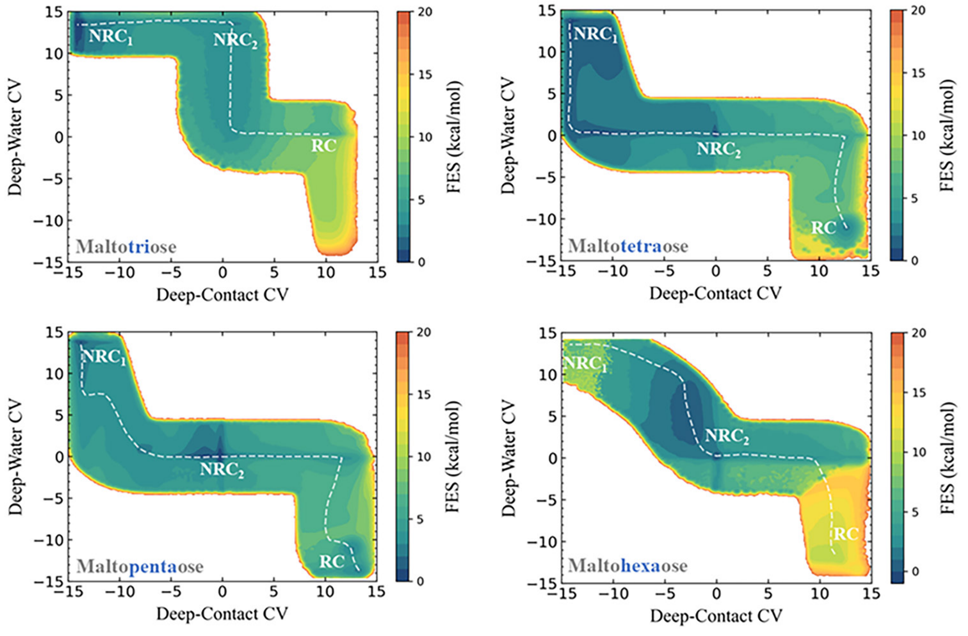

Fig. 3. FES projected into the Deep-Contact and Deep-Water CVs space. From *Left* to *Right* of each plot, the number of reactive contacts between the enzyme and the substrate increases, and from *Top* to *Bottom*, the number of water molecules solvating the reactive contacts decreases. The minimum free energy path connecting the three different conformations of the Michaelis complex is shown with a white dashed line. For clarity, the negative of the Deep-Contact CV is plotted.

maltohexaose. Since maltotriose is the smallest among the substrates and cannot completely occupy the active site pocket of HPA, as confirmed by the multiple minima observed in our Deep-CV space (*SI Appendix*, Fig. S1), it experiences the highest entropic penalty while being trapped within the active site pocket at the NRC_2 state. Thus, the conversion from NRC_1 to NRC_2 results in the largest free energy penalty for the HPA–maltotriose complex, whereas, it is most favorable for the HPA:maltohexaose complex (Fig. 4*A* and *SI Appendix*, Table S1).

During the transition from NRC_2 to RC, desolvation is not the main effect anymore and both the substrate and the solvent molecules inside the catalytic pocket require an internal rearrangement to increase the number of reactive contacts. The NRC_2 → RC conversion is energetically unfavorable for the four enzyme–substrate complexes (Fig. 4*A* and *SI Appendix*, Table S1) as reaction-ready catalytic poses require tight contacts with catalytic residues and fewer water molecules nearby the reactive ring, conditions that are only met in a few properly aligned conformations (*SI Appendix*, Fig. S3). Comparisons between the substrate's and the active site pocket's volume (Fig. 4*B*) and structural analysis (Fig. 4*C*) reveal that maltotetraose aligns perfectly within the active site pocket in the RC state, resulting in the lowest positive $\Delta G_{\mathrm{NRC2 \to RC}}$ (Fig. 4*A* and *SI Appendix*, Table S1). Conversely, for maltohexaose, the glucose units at each end of the substrate largely remain outside the active site cavity in the RC (Fig. 4 *B* and *C*), leading to the highest positive $\Delta G_{\mathrm{NRC2 \to RC}}$.

To correlate the free energetics of the systems under study with their experimental k_{cat}, we computed the probability to be in the RC,

$$\mathrm{NRC}_1 \rightleftharpoons \mathrm{NRC}_2 \rightleftharpoons \mathrm{RC}.$$

Assuming that equilibrium is established between the three states, the probability of reaching RC starting from NRC_1 is $p_{\mathrm{RC}} = \exp(-(\Delta G_{\mathrm{NRC1 \to NRC2}} + \Delta G_{\mathrm{NRC2 \to RC}})/k_B T)$. Using the values of these free energy differences between the successive binding modes from Fig. 4*A* and *SI Appendix*, Table S1, we computed p_{RC} and correlated it with the experimental k_{cat} (Fig. 4*D*). Fig. 4*D* shows a clear correlation between the probability of being in a RC and the catalytic activity. Specifically, when the probability to be in a RC is lowest, as observed with maltotriose, k_{cat} is also lowest. Conversely, when p_{RC} is highest (maltopentaose), k_{cat} increases correspondingly.

In conclusion, our findings show that data-driven ML-CVs, when trained with the appropriate set of descriptors, can be effectively transferred to simulate the dynamics of other systems that exhibit the same binding poses. This approach allowed bypassing the training process of specific CVs for each substrate, which typically requires extensive knowledge about the system and significant user involvement to select meaningful physical descriptors. By using transferable deep-learned CVs, we expedited the correlation of enzyme kinetics across different substrates, streamlining the process significantly.

Although we have tested our approach only on HPA in this study, we expect these data-driven CVs to be transferable to other orthologous enzymes with evolutionarily conserved active sites. However, we may anticipate a potential transferability loss for enzymes with different active sites or for describing different binding modes of substrates with varying scaffolds and chemistry. To overcome this limitation, additional refinement may be required. This refinement step could involve training on a broader dataset covering diverse protein and ligand data and accounting for critical chemical and physical factors such as protein homogeneity and ligand size, shape, and functional groups.

In particular, in the case of CV associated with reactive contact formation, a discovery-based approach (39, 40) could help differentiate

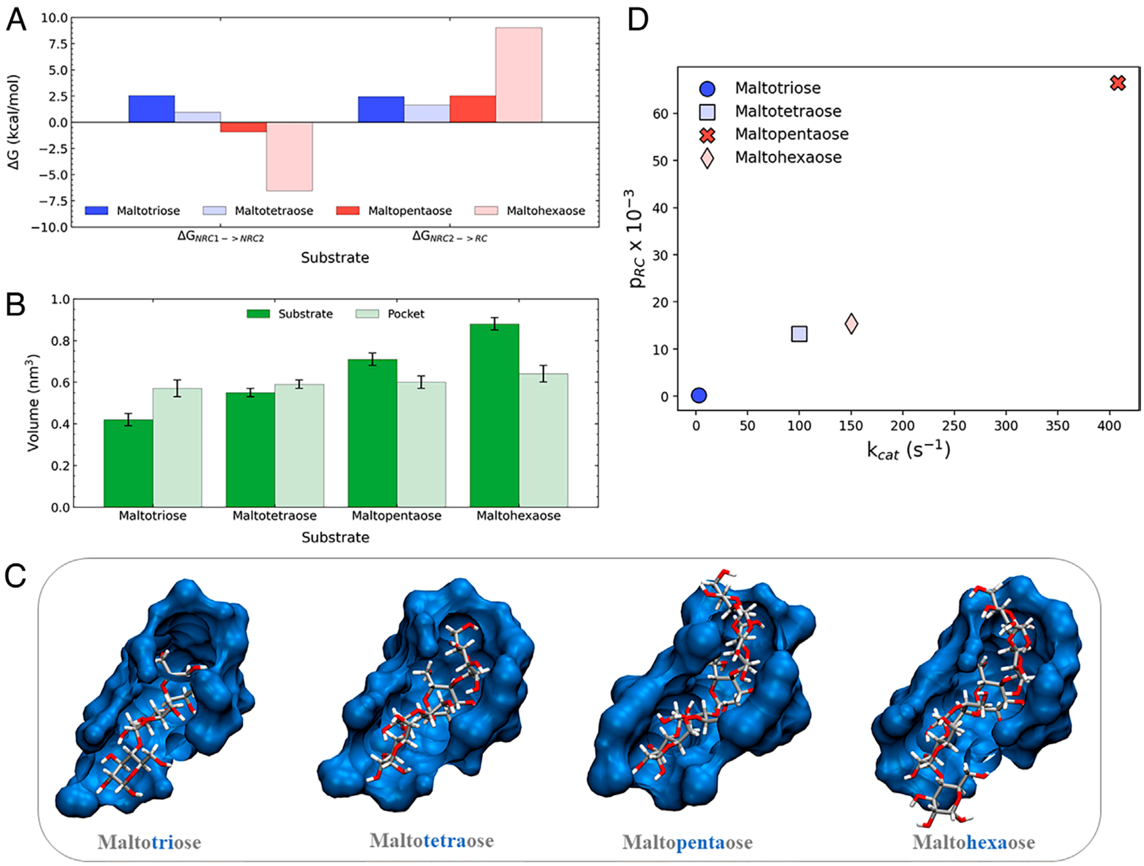

Fig. 4. (*A*) Bar plot for $\Delta G_{\mathrm{NRC1 \to NRC2}}$ and $\Delta G_{\mathrm{NRC2 \to RC}}$ calculated from the corresponding FES as shown in Fig. 3. (*B*) Bar plot comparing the volume of active site pocket with the volume of the substrate in RC. (*C*) Substrate bound to the active site pocket of the enzyme (shown as blue surface) in RC. (*D*) Correlation between p_{RC} and experimental k_{cat}.

reactive conformations of the Michaelis complex from the nonreactive ones (6), allowing for the identification of a set of reactive contacts applicable across all the systems under study. The CV associated with ligand hydration, being more substrate-dependent, may also require retraining for substrates with different chemical properties. For a new system, the transferability of the CVs or the need of retraining could be assessed by running short MD simulations and evaluating the ability of the CVs in promoting back and forth transition between the minima.

Finally, the transferability demonstrated in this study could be highly beneficial in drug discovery, given the crucial role of ligand hydration in the binding process. In this context, the CVs used to compute the binding affinity of a parent drug can be transferred to screen its derivatives, thereby enhancing the design of more effective drugs.

Materials and Methods

Modeling of HPA:Malto-Oligosaccharides Complexes. Different X-ray structures were considered to assemble the initial models, based on the similarity between ligands binding the HPA active site and the malto-oligosaccharides to be modeled. The system encompassing the HPA:maltopentaose was taken from previous work (10), in which the X-ray structure with PDB ID 1CPU (2 Å resolution) including HPA in complex with an acarbose-based pentasaccharide (16) was used. Since no X-ray structure with HPA binding a maltotetraose-like ligand was available, and upon comparison of the binding mode of acarbose-derived penta- and hexasaccharides (*SI Appendix*, Fig. S4), we built the maltotetraose substrate by removing the terminal nonreducing unit of the maltopentaose substrate modeled in the HPA:maltopentaose complex, attending to the product ratio distribution described in the literature (16).

The HPA:maltotriose complex was built from the X-ray structure with PDB ID 1XCW (2 Å resolution) which included HPA in complex with an acarbose-derived trisaccharide (41); the maltotriose substrate was then modeled from the acarbose-derived trisaccharide with PyMOL visualization software. The maltohexaose substrate was built after the acarbose-derived hexasaccharide in complex with HPA in X-ray structure with PDB ID 1XH2 (2.2 Å resolution) (42). Since this HPA structure lacked the calcium and chloride ions required for HPA activity and included a mutation at the Ca^{2+}-binding site (N298S), and the HPA X-ray structures overlapped between PDB IDs 1CPU and 1XH2 (root-mean-square deviation of 0.14 Å over 3,379 atoms), we have chosen to model the maltohexaose substrate from the acarbose-derived hexasaccharide from PDB ID 1XH2 into the active site of HPA from PDB ID 1CPU.

Glu233 was maintained in a neutral state because it acts as an acid in the catalytic process. All other titratable residues were maintained in their standard protonation state at pH 7. All solute complexes were then enclosed in a rectangular box with a 12 Å minimum distance from the limits of the solute complex. Sodium counterions were added to neutralize the charge of the system.

Molecular Dynamics Simulations. The HPA was described with FF99SB (43) and the malto-oligosaccharide substrates were described with the GLYCAM-06H (44). All waters were described with the TIP3P model (45). All topologies and initial coordinates are supplied in the PLUMED-NEST repository.

We applied the following simulation protocol to each substrate:

a) A steepest descent energy minimization was performed on the system, maintaining position restraints on heavy atoms with a force constant of 10^3 kJ/mol/rad^2.
b) The minimized structure was initially equilibrated for 200 ps in the isothermal-isovolume (NVT) ensemble (T = 300 K) followed by a 200 ps simulation in the isothermal-isobaric ensemble (NPT) (T = 300 K, $P = 1$ bar) simulation, both with position restraints.
c) Subsequently, a 1 ns NPT simulation was conducted with the force constant for the position restraints reduced to 10 kJ/mol/rad^2. This was followed by another 10 ns NPT simulation, during which all position restraints were removed to equilibrate the volume of the simulation box.
d) Then, a 1 μs NVT production run was carried out without position restraints.
e) Finally, four independent unbiased MD simulations were initiated from four distinct starting configurations, each randomly selected from the initial MD production run (point d). In total, five independent unbiased MD simulations were performed, which were used to generate the density plots shown in *SI Appendix*, Figs. S1 and S2.

The Bussi–Donadio–Parrinello velocity rescaling thermostat (46) was employed with a coupling constant of 0.1 for equilibration and 1.0 for production at 300 K in the NVT runs. For the NPT runs, the same thermostat with a coupling constant of 1.0 ps was used to equilibrate the temperature at 300 K, while the Parrinello–Rahman barostat (47, 48), with a coupling constant of 1.0 ps, was used to maintain the pressure at 1 bar. An integration time step of 2 fs was utilized, along with the linear constraint solver (LINCS) (49) constraints on all bonds involving hydrogen atoms. The particle mesh Ewald method (50), with a cutoff distance of 10 Å, was applied to handle long-range electrostatic interactions. All simulations were conducted using GROMACS 2021.5 (51), with trajectory frames saved every 10 ps.

Free Energy Calculations.
Data-driven CVs. To calculate the FES for the transformation from NRC to RC of maltotriose, maltotetraose, and maltohexaose, we used Deep-Contact and Deep-Water CV, trained on maltopentaose using Deep-TDA (38). Extensive details about the training of these CVs can be found in our previous work (6).

Briefly, Deep-TDA uses physical descriptors (i.e., distances, angles) as inputs to a neural network, which is optimized to map these descriptors into a low-dimensional CV space. This optimization aligns the distribution of training data with predefined Gaussian targets, ensuring that configurations from different metastable states match these targets in the CV space.

The selection of input descriptors for the neural network is essential in this context. We developed two CVs to characterize the reactive contacts between the substrate and the enzyme (Deep-Contact), as well as the arrangement of water molecules within the enzymatic active site cavity (Deep-Water).

Deep-Contact CV considers the reactive contacts of the substrate within the catalytic pocket. During its training, the neural network was provided with structural parameters that describe reactive criteria (*SI Appendix*, Fig. S5A), including the availability of nucleophiles, proton donors, and structural hydrogen bonds. Twelve distances and one torsional angle (as shown in *SI Appendix*, Fig. S5B) were used as descriptors.
Specifically, the geometric descriptors include:

1) Distances between O1@Asp197, O2@Asp197, and the C_1 atom of the reactive sugar ring ($\mathbf{d_1}$ and $\mathbf{d_2}$), hydrogen bonds between the $-CH_2OH$ of the sugar with the nucleophiles ($\mathbf{d_3}$ and $\mathbf{d_4}$), and hydrogen bonds between Glu233 and the nucleophiles ($\mathbf{d_5}$ and $\mathbf{d_6}$). These descriptors define the availability of nucleophile.
2) Distance between the glycosidic oxygen O_{gly} and the carboxylic -OH group of Glu233 ($\mathbf{d_7}$). This descriptor takes into account the availability of the proton donor.
3) Hydrogen bonds of the hydroxyl groups O2-H and O3-H of the reactive sugar ring with the carboxylic oxygens of Asp300 ($\mathbf{d_8}$ to $\mathbf{d_{11}}$), side chain torsional angle N-C_α-C_β-C_γ of Asp300 ($\boldsymbol{\chi_1}$). These parameters describe structural hydrogen bonds.
4) Z-component of the nucleophile-ligand distance $\mathbf{d_{12}}$ (distance of the center of mass of the reactive sugar ring from the C_α of Asp197) is also included as an input descriptor.

Deep-Water CV was trained to describe the water arrangement in the three binding modes. We used the coordination number of water molecules around strategic binding positions as input descriptors for the neural network. Sixteen descriptors were used in the training (*SI Appendix*, Fig. S5C). The first 10 descriptors represented the water coordination around the ten hydration spots (where water molecules reside 0.5 ns or beyond), identified using the approach of Ansari et al. (52) and described in detail in our previous work (6). To describe the solvation of the ligand, we introduced descriptors 11 to 14, which were the coordination numbers of water molecules around the polar groups of the reactive sugar ring (O2, O3, O_{gly}, O6, in *SI Appendix*, Fig. S5C). The solvation of the binding pocket was captured by descriptors 15 and 16, representing the coordination numbers between water and the carbon atoms of the carboxylate groups of Asp197 and Asp300, respectively. The solvation of the carboxylate group of Glu233 was not considered, as it remained solvated by W_1 in all three states and thus could not differentiate these states (Fig. 2 *B* and *C*).

Path CV. Before recovering the FES, we projected all the unbiased classical MD trajectories into the two-dimensional Deep-Contact-Deep-Water space (*SI Appendix*, Fig. S1). In this two-dimensional (2D) space, the three states NRC_1, NRC_2, and RC were clearly distinguished. We identified the transition path from state NRC_1 to state RC via state NRC_2 (shown in white dashed line in *SI Appendix*, Fig. S1). Considering this path, we designed a Path CV to enable more efficient sampling along the path while ignoring irrelevant parts of the FES. This approach allowed for faster convergence. The Path CV, developed by Branduardi et al. (53), enables computation of progress along a high-dimensional path and the distance from that path. The progress along the path(s) is given by:

$$s = \frac{\sum_{i=1}^{N} i e^{(-\lambda R[X-X_i])}}{\sum_{i=1}^{N} e^{(-\lambda R[X-X_i])}},$$

while the distance from the path (z) is calculated as:

$$z = -\frac{1}{\lambda}\ln\left[\sum_{i=1}^{N} e^{\left(-\lambda R\left[X-X_i\right]\right)}\right].$$

Here, N high-dimensional frames (X_i) are used to describe the path in the high-dimensional space. Hence, s and z are functions of the distances from each of the high-dimensional frames $R[X-X_i]$. In our calculations, we placed 37 equidistant state points ($\{X_i\}$) (using *pathtools* module implemented in PLUMED) to construct the reference path. Herein, $\{R[X-X_i]\}$ represents Euclidean distances of a dynamic state point X during the evolution of the path from each of the reference state points $\{X_i\}$ in the 2D space of normalized Deep-Contact and Deep-Water CVs. The λ values are set to be 200 au, which are inversely proportional to the Euclidean distance between two consecutive states on the reference path. To limit the sampling only around the reference path, we have included an upper wall restraint at $z = 0.01$ au with a force constant of 5×10^5 au.

Enhanced Sampling for estimating fes using OPES with path CV. To calculate the FES for the transformation from NRC to RC of HPA–maltotriose, HPA–maltotetraose, and HPA–maltohexaose, we enhanced the fluctuation of the Path CV via the OPES method (20). To estimate the FES, we used OPES with four multiple walkers, setting the BARRIER to 13 kcal/mol and PACE to 500. Initially, an equilibration phase was conducted where all walkers contributed to the bias potential, reaching a quasistatic regime within 250 ns of simulation for each walker. Subsequently, we decoupled the walkers and continued their simulations independently up to 1 μs for each walker, using the previously generated external potential as a bias (*SI Appendix*, Fig. S6). These trajectories were then merged into a 3 μs long single trajectory, which was used to estimate the FES and assess its convergence. Same protocol has been used for HPA-maltopentaose in the work of Das et al. (6)

We reconstructed the FES from reweighting (20) and performed block analysis to test its convergence. For numerical stability, we filter out the outlier configurations (less than 1% of the total) where the bias deposited on the walls over the z-component of nucleophile-substrate distance, deep CVs, and the path CV are larger than a threshold of 2.4 kcal/mol. We split the resulting trajectory into $N_B = 10$ blocks of ≈300 ns each. In each block i, we independently evaluate the one-dimensional (1D) FES along a deep CV (Deep-Contact CV or Deep-Water CV). Every block i has a different statistical weight measured as

$$w_i = \sum_j e^{\beta V(j)},$$

where index j runs over the configurations in the block and V is the bias potential. The effective sample size

$$N_{eff} = \frac{\left(\sum_{i=1}^{N_B} w_i\right)^2}{\sum_{i=1}^{N_B} w_i^{\,2}},$$

measures the quality of the weighted block average with respect to the weight distribution in the blocks. In the ideal case of low correlation, $N_{eff} \to N_B$. We observe a Neff ≈9.5 with an average FES uncertainty (measured as weighted SE to 1D FES between the blocks) of 0.1 kcal/mol for all the HPA-substrate systems, except for HPA–maltotriose which shows an average FES uncertainty of 0.5 kcal/mol (*SI Appendix*, Figs. S7–S9).

Volume Analysis. All enzymatic residues within 3 Å of the substrate were considered to construct the active site cavity. The volume of this cavity (as shown in Fig. 4*B*) was computed by building a convex hull (using α-shape toolbox in python) incorporating all the selected residues. All the atoms of corresponding substrate were considered to calculate its volume using similar protocol as mentioned before.

Data, Materials, and Software Availability. Input files data have been deposited in PLUMED NEST repository with plumID: 24.025 (54).

ACKNOWLEDGMENTS. We would like to thank Enrico Trizio, Dhiman Ray and Axel Tosello Gardini for helpful discussions. R.P.P.N. and M.J.R. acknowledge the support from Fundação para a Ciência e Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior (FCT/MCTES): LA/P/0008/2020 DOI: 10.54499/LA/P/0008/2020, UIDP/50006/2020 DOI: 10.54499/UIDP/50006/2020, and UIDB/50006/2020 DOI: 10.54499/UIDB/50006/2020). R.P.P.N. further thanks FCT for funding through the Individual Call to Scientific Employment Stimulus (Ref. 2021.00391.CEECIND/CP1662/CT0003 DOI: 10.54499/2021.00391.CEECIND/CP1662/CT0003).

Author affiliations: [a]Atomistic Simulation Research Line, Italian Institute of Technology, Genova GE 16152, Italy; and [b]Laboratório Associado para a Química Verde, Rede de Química e Tecnologia, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal

1. P. Hanoian, C. T. Liu, S. Hammes-Schiffer, S. Benkovic, Perspectives on electrostatics and conformational motions in enzyme catalysis. *Acc. Chem. Res.* **48**, 482–489 (2015).
2. R. Callender, R. B. Dyer, The dynamical nature of enzymatic catalysis. *Acc. Chem. Res.* **48**, 407–413 (2015).
3. J. Deng, Q. Cui, Second-shell residues contribute to catalysis by predominately preorganizing the apo state in PafA. *J. Am. Chem. Soc.* **145**, 11333–11347 (2023).
4. A. T. Bogetti, M. C. Zwier, L. T. Chong, Revisiting textbook azide-clock reactions: A "propeller-crawling" mechanism explains differences in rates. *J. Am. Chem. Soc.* **146**, 12828–12835 (2024).
5. S. Romero-Téllez, A. Cruz, L. Masgrau, À. González-Lafont, J. M. Lluch, Accounting for the instantaneous disorder in the enzyme–substrate Michaelis complex to calculate the Gibbs free energy barrier of an enzyme reaction. *Phys. Chem. Chem. Phys.* **23**, 13042–13054 (2021).
6. S. Das, U. Raucci, R. P. P. Neves, M. J. Ramos, M. Parrinello, How and when does an enzyme react? Unraveling α-Amylase catalytic activity with enhanced sampling techniques. *ACS Catal.* **13**, 8092–8098 (2023).
7. D. Ray, S. Das, U. Raucci, Kinetic view of enzyme catalysis from enhanced sampling QM/MM simulations. *J. Chem. Inf. Model.* **64**, 3953–3958 (2024).
8. S. Hur, T. C. Bruice, The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12015–12020 (2003), 10.1073/pnas.1534873100.
9. N. F. Brás, P. A. Fernandes, M. J. Ramos, QM/MM studies on the β-Galactosidase catalytic mechanism: Hydrolysis and transglycosylation reactions. *J. Chem. Theory Comput.* **6**, 421–433 (2010), 10.1021/ct900530f.
10. D. Santos-Martins, A. R. Calixto, P. A. Fernandes, M. J. Ramos, A buried water molecule influences reactivity in α-Amylase on a subnanosecond time scale. *ACS Catal.* **8**, 4055–4063 (2018), 10.1021/acscatal.7b04400.
11. L. Pfaff *et al.*, Multiple substrate binding mode-guided engineering of a thermophilic PET hydrolase. *ACS Catal.* **12**, 9790–9800 (2022), 10.1021/acscatal.2c02275.
12. D. A. Pomeranz Krummel, S. Altman, Multiple binding modes of substrate to the catalytic RNA subunit of RNase P from Escherichia coli. *RNA* **5**, 1021–1033 (1999), 10.1017/s1355838299990416.
13. C. P. Wong *et al.*, Two distinct substrate binding modes for the normal and reverse prenylation of hapalindoles by the prenyltransferase AmbP3. *Angew. Chem. Int. Ed. Engl.* **57**, 560–563 (2018), 10.1002/anie.201710682.
14. N. R. Wong, R. Sundar, S. Kazanis, J. Biswas, T. C. Pochapsky, Conformational heterogeneity suggests multiple substrate binding modes in CYP106A2. *J. Inorg. Biochem.* **241**, 112129 (2023), 10.1016/j.jinorgbio.2023.112129.
15. K. S. Bak-Jensen *et al.*, Tyrosine 105 and threonine 212 at outermost substrate binding subsites −6 and +4 control substrate specificity, oligosaccharide cleavage patterns, and multiple binding modes of barley α-Amylase 1. *J. Biol. Chem.* **279**, 10093–10102 (2004), 10.1074/jbc.M312825200.
16. G. D. Brayer *et al.*, Subsite mapping of the human pancreatic α-amylase active site through structural, kinetic, and mutagenesis techniques. *Biochemistry* **39**, 4778–4791 (2000), 10.1021/bi9921182.
17. M. A. Maria-Solano, E. Serrano-Hervás, A. Romero-Rivera, J. Iglesias-Fernández, S. Osuna, Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun.* **54**, 6622–6634 (2018).

18. S. Jayaraj, S. Suresh, R. K. Kadeppagari, Amylase inhibitors and their biomedical applications. *Starch-Stärke* **65**, 535–542 (2013).
19. H. Oliveira *et al.*, Anthocyanins as antidiabetic agents—in vitro and in silico approaches of preventive and therapeutic effects. *Molecules* **25**, 3813 (2020).
20. M. Invernizzi, M. Parrinello, Rethinking metadynamics: From bias potentials to probability distributions. *J. Phys. Chem. Lett.* **11**, 2731–2736 (2020), 10.1021/acs.jpclett.0c00497.
21. M. Invernizzi, P. M. Piaggi, M. Parrinello, Unified approach to enhanced sampling. *Phys. Rev. X* **10**, 041034 (2020), 10.1103/PhysRevX.10.041034.
22. W. Chen, A. L. Ferguson, Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **39**, 2079–2102 (2018).
23. M. Chen, Collective variable-based enhanced sampling and machine learning. *Eur. Phys. J. B* **94**, 1–17 (2021).
24. D. Mendels, G. Piccini, M. Parrinello, Collective variables from local fluctuations. *J. Phys. Chem. Lett.* **9**, 2776–2781 (2018).
25. L. Bonati, V. Rizzi, M. Parrinello, Data-driven collective variables for enhanced sampling. *J. Phys. Chem. Lett.* **11**, 2998–3004 (2020).
26. G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, F. Noé, Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 014101 (2013).
27. L. Bonati, G. Piccini, M. Parrinello, Deep learning the slow modes for rare events sampling. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2113533118 (2021).
28. M. M. Sultan, V. S. Pande, Automated design of collective variables using supervised machine learning. *J. Chem. Phys.* **149**, 094103 (2018).
29. J. M. L. Ribeiro, P. Bravo, Y. Wang, P. Tiwary, Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **149**, 074105 (2018).
30. L. Sun *et al.*, Multitask machine learning of collective variables for enhanced sampling of rare events. *J. Chem. Theory Comput.* **18**, 2341–2353 (2022).
31. F. Hooft, A. Pérez de Alba Ortiz, B. Ensing, Discovering collective variables of molecular transitions via genetic algorithms and neural networks. *J. Chem. Theory Comput.* **17**, 2294–2306 (2021).
32. D. Ray, E. Trizio, M. Parrinello, Deep learning collective variables from transition path ensemble. *J. Chem. Phys.* **158**, 204101 (2023).
33. M. A. Rohrdanz, W. Zheng, C. Clementi, Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.* **64**, 295–316 (2013).
34. X. Qin *et al.*, Structures of human pancreatic α-amylase in complex with acarviostatins: Implications for drug design against type II diabetes. *J. Struct. Biol.* **174**, 196–202 (2011).
35. G. P. Pinto *et al.*, Establishing the catalytic mechanism of human pancreatic α-amylase with QM/MM methods. *J. Chem. Theory Comput.* **11**, 2508–2516 (2015), 10.1021/acs.jctc.5b00222.
36. R. P. P. Neves, P. A. Fernandes, M. J. Ramos, Role of enzyme and active site conformational dynamics in the catalysis by α-amylase explored with QM/MM molecular dynamics. *J. Chem. Inf. Model.* **62**, 3638–3650 (2022), 10.1021/acs.jcim.2c00691.
37. R. P. Neves, A. V. Cunha, P. A. Fernandes, M. J. Ramos, Towards the accurate thermodynamic characterization of enzyme reaction mechanisms. *ChemPhysChem* **23**, e202200159 (2022).
38. E. Trizio, M. Parrinello, From enhanced sampling to reaction profiles. *J. Phys. Chem. Lett.* **12**, 8621–8626 (2021), 10.1021/acs.jpclett.1c02317.
39. F. Pietrucci, W. Andreoni, Graph theory meets ab initio molecular dynamics: Atomic structures and transformations at the nanoscale. *Phys. Rev. Lett.* **107**, 085504 (2011), 10.1103/PhysRevLett.107.085504.
40. U. Raucci, V. Rizzi, M. Parrinello, Discover, sample, and refine: Exploring chemistry with enhanced sampling techniques. *J. Phys. Chem. Lett.* **13**, 1424–1430 (2022), 10.1021/acs.jpclett.1c03993.
41. C. Li *et al.*, Acarbose rearrangement mechanism implied by the kinetic and structural analysis of human pancreatic α-amylase in complex with analogues and their elongated counterparts. *Biochemistry* **44**, 3347–3357 (2005).
42. R. Maurus *et al.*, Structural and mechanistic studies of chloride induced activation of human pancreatic α-amylase. *Protein Sci.* **14**, 743–755 (2005).
43. V. Hornak *et al.*, Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).
44. K. N. Kirschner *et al.*, GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622–655 (2008).
45. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983), 10.1063/1.445869.
46. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007), 10.1063/1.2408420.
47. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981), 10.1063/1.328693.
48. S. Nosé, M. L. Klein, Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **50**, 1055–1076 (1983), 10.1080/00268978300102851.
49. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997), 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
50. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993), 10.1063/1.464397.
51. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015), 10.1016/j.softx.2015.06.001.
52. N. Ansari, V. Rizzi, M. Parrinello, Water regulates the residence time of benzamidine in trypsin. *Nat. Commun.* **13**, 5438 (2022), 10.1038/s41467-022-33104-3.
53. D. Branduardi, F. L. Gervasio, M. Parrinello, From A to B in free energy space. *J. Chem. Phys.* **126**, 054103 (2007).
54. S. Das, Correlating enzymatic reactivity for different substrates using transferable data-driven collective variables. PLUMED NEST. https://www.plumed-nest.org/eggs/24/025/. Deposited 26 October 2024.