scientific reports

Check for updates

OPEN

Geometry-encoded molecular dynamics enables deep learning insights into P450 regiospecificity control

Denis Pompon ✉, Luis F. Garcia-Alles & Philippe Urban

Cytochrome P450 1A2, as many isoenzymes, can generate multiple metabolites from a single substrate. A loose coupling between substrate binding and oxygen activation makes possible substrate reorientations at the active site prior to catalysis. In the present work, caffeine oxidation to alternative bioactive compounds was used to decipher this pluripotency. A model involving two interacting subsites capable of sequentially accommodating one or two caffeine molecules was considered. Molecular dynamics was used to characterize subsite interactions and feed a dedicated geometric encoding of trajectories that was coupled to dimensional reductions and differential machine learning. The two subsites differentially control caffeine orientations and can exchange substrate through a phenylalanine gated mechanism. This exchange can be locked by the presence of a second bound molecule. Complementary roles of subsites in progressively determining the caffeine orientation during its approach to active oxygen were examined. Interestingly, substrate face flipping becomes impaired upon entry into the rather flat active site. This makes the mechanisms that define the orientation of caffeine relative to active oxygen dependent on the substrate face oriented toward heme. Globally, this evidenced that P450 1A2 regioselectivity results from local determinants combined with subsite interactions and caffeine face preselection at a longer distance.

Keywords Deep learning, Molecular dynamic, Geometric encoding, P450, Caffeine, Regioselectivity

Molecular dynamics (MD) has been frequently used to analyze protein substrate interactions during catalytic processes, predict consequences of natural polymorphism, or design enzymes endowed with optimized or even new functions of interest[1]. Among the questions addressed are mechanisms capable of selecting a particular orientation of the substrate at an active site, potentially leading to the formation of alternate products. A specific case occurs with cytochrome P450, for which the lifetime of enzyme-bound activated oxygen can be long enough compared to the repositioning or release of the substrate from the active site. This case frequently results in the formation of multiple products and/or partially uncoupled reactions[2]. Density of presence of oxidizable substrate atoms at suitable distances from activated oxygen combines with their intrinsic chemical reactivities to define product formation[3].

In the specific case of human P450 1A2 and caffeine, at least four different metabolites are formed in unequal proportions[4]. The significant difference between the product profiles reported at low and high substrate concentrations suggested that several caffeine molecules could simultaneously interact with the enzyme[5]. Observation of cooperative behaviors when different substrates were simultaneously added further supports this finding[6–8]. Interestingly, similar metabolite patterns were observed when P450 1A2 activity was supported by the regular electron transfer chain leading to oxygen activation or by an active oxygen donor added as a co-substrate[9]. This implies that the latter and caffeine must sequentially bind to the active site and exchange during the lifetime of the formed P450 ferryl-oxo species. In silico docking and access channel predictions[10] extended this view by defining potential caffeine subsites and several MD studies addressed in more detail their caffeine binding modes[4,11]. However, the analysis of predicted substrate trajectories appeared to be only partially capable of accounting for the experimental regiospecificity and remained mostly descriptive. The interactions of water molecules[4], long-distance conformational effects[11,12], the lipid environment[3] and the intrinsic reactivity of caffeine atoms[13] were multiple factors that made a rationale difficult to establish.

Toulouse Biotechnology Institute, Université de Toulouse, CNRS, INRAE, INSA, 135 Avenue de Rangueil, Toulouse, France. ✉email: dpompon@insa-toulouse.fr

Machine learning approaches are particularly prone to aggregate in a unique model such multifactorial mechanisms. Placed upstream or downstream of MD, they facilitated data integration and helped to develop predictive models for various enzyme systems[14–16]. A critical factor remained the encoding of MD trajectories in a way that reduces the number of degrees of freedom and facilitates the exploration of the underlying mechanisms. Therefore, pretrained protein languages were developed and used to predict P450 substrates or inhibitors[17,18]. Recently, coupling of geometrical encoding of protein structures with MD has appeared of particular interest[19–22]. However, while facilitating machine learning processes, geometric approaches were frequently too generalist to permit easy mechanistic interpretation. Here, our objective was to address this question, taking as an example the determinism of regioselectivity for caffeine metabolism by human P450 1A2. A geometric model of the active site was built and used to define a set of dedicated geometrical descriptors fed by MD trajectories. The network of relationships evidenced between them enables us to use dimensional reduction and differential machine learning approaches to compare mechanisms of substrate orientation in alternate contexts including substrate face orientation and single or multiple occupancy at active site.

Results

Two caffeine molecules can bind simultaneously at the active site of P450 1A2

Caffeine, an almost planar molecule, is actively metabolized by P450 1A2 at the level of its C12, C10 and C14 methyl groups giving rise respectively to formation of paraxanthine (85%), theobromine, (10%) and theophylline (5%) and of minor amounts of C8 hydroxylation. Substrate docking (HADDOCK), access channel analysis (CAVER) and MD approaches offer complementary views of caffeine binding modes leading to product formation or playing an effector role. Figure 1 (panels a, c) illustrates that these approaches defined partially overlapping regions delimiting two protein subsites named producer and effector sites hereafter. These subsites take place on the distal face of the heme opposite to the interface formed between the P450 and the FMN domain the reductase (Fig. 1b). The two subsites are linked by a rather coplanar protein envelope, permissive for substrate rotation on its own plane, but excluding its face flipping. The producer subsite appears relatively amphiphilic compared to the more apolar effector subsite that particularly involves two phenylalanine residues frequently involved in substrate stacking and a last one closing the pocket (Fig. S1). Although caffeine does not have an asymmetric center, the substrate face oriented toward heme (designed as R or L configurations in this paper) was frozen at the active site, leading to pseudo-isomers of the substrate in the protein context. The caffeine binding energies at the two subsites were predicted to be fairly similar and consistent with an experimental caffeine affinity in the range of 10 to 100 micromolar. Figure 2 defines nomenclature for the caffeine side facing the heme, atom naming and orientation codes used throughout the paper in consistency with acquired MD data and R-scripts used for analysis. At producer site, at least one of the caffeine methyl groups is 3–5 Å distant from the ferryl-oxo oxygen, enabling metabolite formation. This distance increases to 8–12 Å at effector site, thus precluding any metabolite formation. Frequent substrate stacking with Phe-202 (226) (see Methods for atom numbering) and more marginally Phe-236 (260) also characterized the effector site. During MD courses, the caffeine trajectories observed at the two subsites were very marginally overlapping, thus permitting a double occupancy without significant substrate contacts.

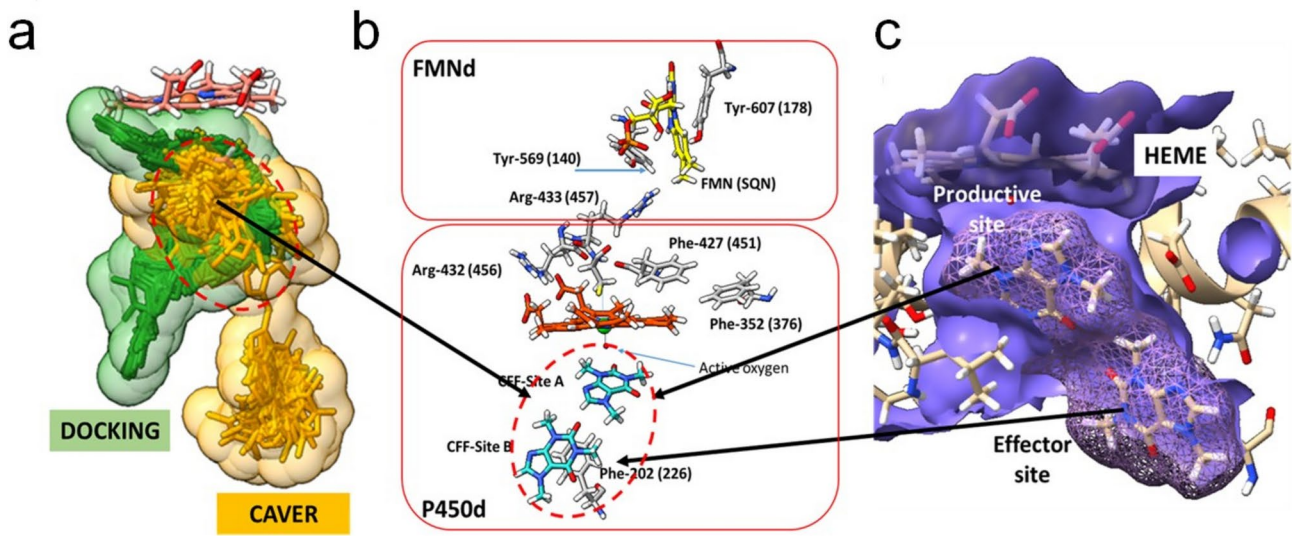

Fig. 1. Docking, CAVER channel, and subsite definition for caffeine bound to P450 1A2. (**a**) HADDOCK docking poses (yellow), and CAVER active site access trajectories (green) for caffeine binding. (**b**) Topologies of the producer and effector caffeine binding sites relative to heme and to the P450 bound FMN domain of the reductase. (**c**) Details of the protein envelope at the producer and effector sites. Phe226 (260), and Phe202 (226) were represented in panel b and heme on tops. Amino-acid numbering for the corresponding full-length proteins are given into brackets.

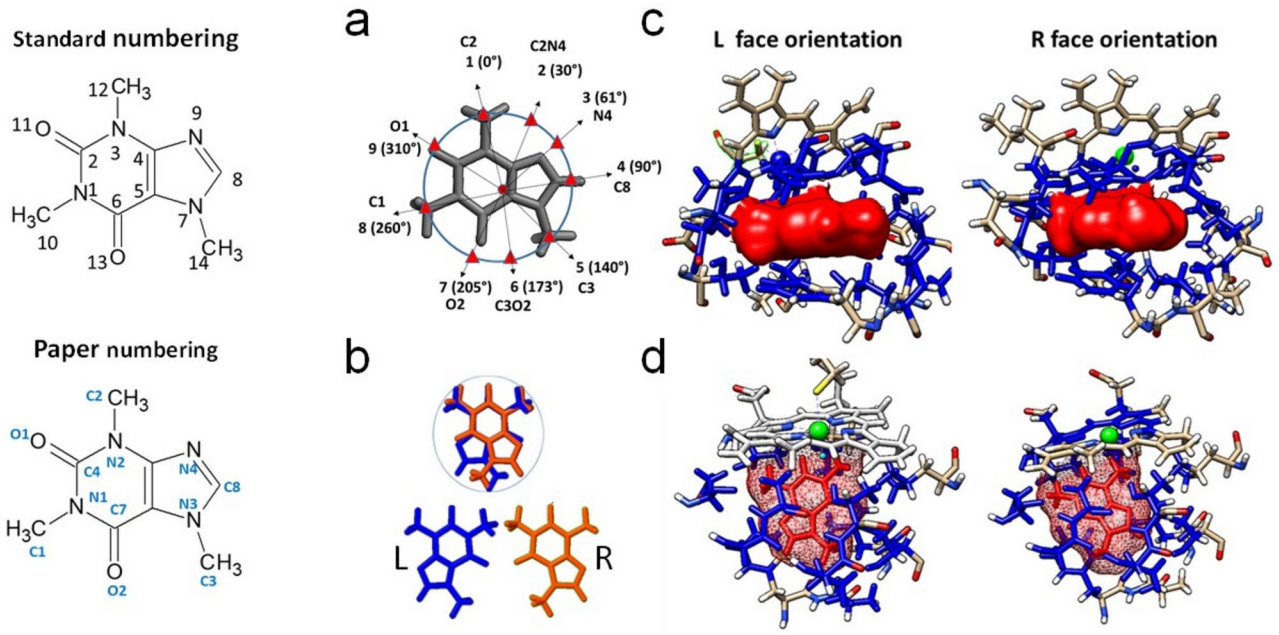

Fig. 2. Atom numbering, definition of face configurations, and orientation codes for caffeine. On the left, the standard and the corresponding numbering used in this paper for caffeine atoms are given. (**a**) definition of A2 angles (0–310°), numerical indexes (1–9) and orientation codes (C2-O1), respectively, characterizing caffeine orientations relative to active oxygen projection on the caffeine plane. (**b**) Views of caffeine faces (L or R configuration) relative to heme observed from the caffeine side opposite to heme. (**c**) views perpendicular to the heme (on top) of caffeine (in red) embedded in active site residues (blue), and (**d**) the same views parallel to the heme (iron in green).

Molecular dynamics settings and caffeine orientation encoding

The modeled complex previously described[23] between the P450-reducase (restricted to its FMN domain) and the human P450 1A2 was used as template to set up a MD model in explicit solvent. This model involved the FMN cofactor in a neutral semiquinone form and the heme in the ferryl-oxo redox state. Such a combination was expected to optimally mimic the conditions prevailing during catalytic events. MD trajectories were initiated using alternate substrate settings inspired by the main observed docking occurrences and sampled at 40 ps intervals. Sixty-two independent dynamics (up to 60 ns each) were used to sample substrate trajectories either for a single or a pair of bound caffeine. In addition, four sets of MD runs, each involving twelve rationally designed caffeine settings (rotamers), were performed. For that, the centroid of caffeine methyl groups (center of the circle that encompasses their carbon atoms) and the caffeine plane were kept unchanged, while the substrate orientation was progressively changed by 30° steps from 0 to 330° (see Sect "Methods"). As a full exploration of the 12×12 possible angle combinations involving the two subsites was not manageable, only the more stable orientation at the producer site was combined with the full set of rotamers at the effector site and reciprocally. Molecular dynamics trajectories were used as a starting material to initiate complementary analysis based on a geometrical representation of the heme-substrate subsystem as summarized in Fig. S2.

A local geometrical model describes the heme-caffeine subsystem

A geometric model (Fig. 3, and Tables S1, S2) was designed to encode relative positions and orientations of heme (including active oxygen) and caffeine. Modelling took advantage of the protein-bound heme cofactor to build a local coordinate system defined by the plane and orientations of porphyrin nitrogens, the iron and the attached active oxygen (Fig. S3). The relative geometry of substrate was described by its own plane, as well as the position of its centroid and orientation of its methyl groups relative to active oxygen. The corresponding set of geometric descriptors was designed to constitute a bijective encoding of Cartesian coordinates while being more easily interpretable in terms of catalytic mechanism. A set of secondary descriptors was derived (Tables S1, S3) and also additionally defined the relative geometries between caffeine and two phenylalanine residues (Phe202 and 226) frequently involved in substrate stacking. Globally, position and orientation of caffeine are characterized by four angles and two distances. Three of them (A0, A1, D2) define the relative positions of the heme and caffeine planes, while three others (A2, A3, D4) define the position and orientation of substrate on its own plane. Noticeably, the A2 angle and the derived numerical index Imin characterize the orientation of caffeine atoms relative to the orthogonal projection of active oxygen in the substrate plane and Dmin the corresponding distance. This combination of distance and angular descriptors can be interpreted as a major determinant of the regioselectivity of reaction.

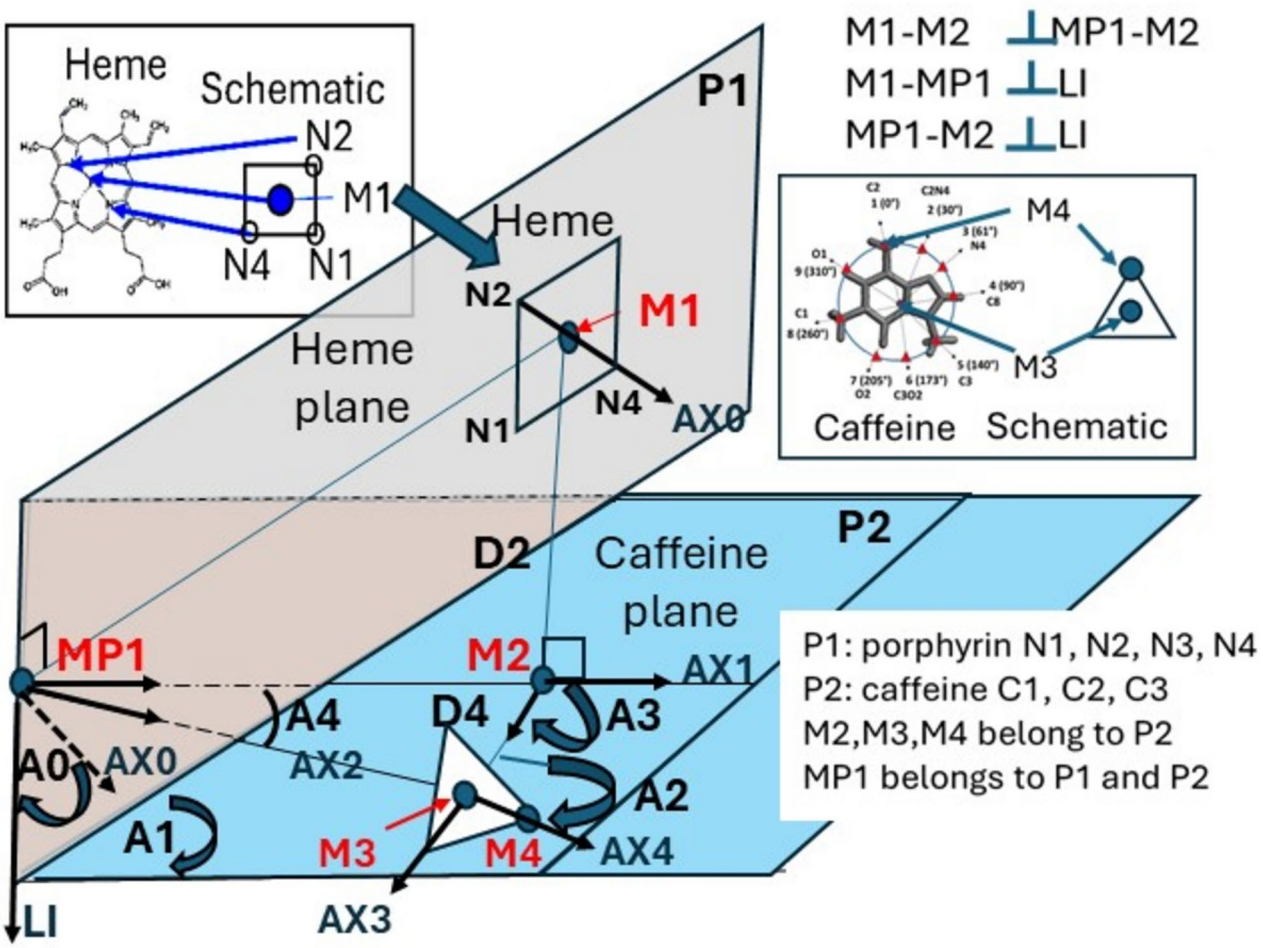

Fig. 3. Definition of geometric descriptors. The heme plane (P1) forms a dihedral angle (A1) with the caffeine plane (P2). M1 is heme iron or active oxygen, depending on the case. M2 is the orthogonal projection of M1 on plane P2. LI is the director vector of the P1-P2 intersect line. MP1 is the common orthogonal projection of M1, and M2 on the line defined by the LI vector. A0 is the angle between the porphyrin N2-N4 axis and the LI vector. M3 is the centroid of the caffeine methyl groups, and M4 is the caffeine C2 (C12) atom. Vector AX1, AX2, AX3, AX4, angles A2, A3, A4 as well as distances D2, D4 are as indicated on figure. Porphyrin N1, N2, N3 and caffeine C1, C2, C3 rotations are used as a reference for A0 and A2, A3, A4 signs, respectively. A1 is in the range of 0 to 180° and D4 is set positive only when M2 and M4 belong to the same side of the P1-P2 dihedra (A1 < 90°).

Statistical encoding of geometrical parameter time-series and relations during MD courses

Observation that the values of geometrical descriptors encoding a given MD trajectory formed sets of Gaussian distributions (Fig. S4) led to an additional degree of simplification. The Gaussian factorization of each descriptor was performed considering values independently of their order in time-series. Major, secondary, and eventually minor Gaussian contributions to each descriptor and trajectory were defined on the basis of the number of MD poses observed. Corresponding sets of means (points) and standard deviations (bars) were plotted for each MD trajectory (labeled on the x-axis) and color-encoded by their levels of contribution in Fig. S5. A clear relationship was evidenced between the A2 angle (or Imin index) that defines caffeine orientation and the Dmin distance (Fig. 4 and Fig. S5). This relationship was expected as a result of local interactions between the substrate and its protein environment in close proximity to active oxygen. However, caffeine orientation and the D4 distance that characterize the proximity between caffeine centroid and the position on caffeine plane that is closest to active oxygen appeared also to be significantly related (Fig. S5). This evidenced that short distance interactions in immediate proximity of active oxygen were involved in the adjustment of caffeine orientations. The progressive orientation of the substrate during its approach to active oxygen was also illustrated by the counts of MD poses exhibiting a specific substrate orientation as a function of the distance to oxygen (Fig. 4 panels c, d). The major predicted orientation in close proximity to oxygen clearly matches with experimental formation of the major metabolite: paraxanthine. In contrast, no correlation was found between orientation and A1 or A2 angles and only a fuzzy one with A0 angles indicating a rather limited effect on orientation of the relative angles of caffeine and heme planes.

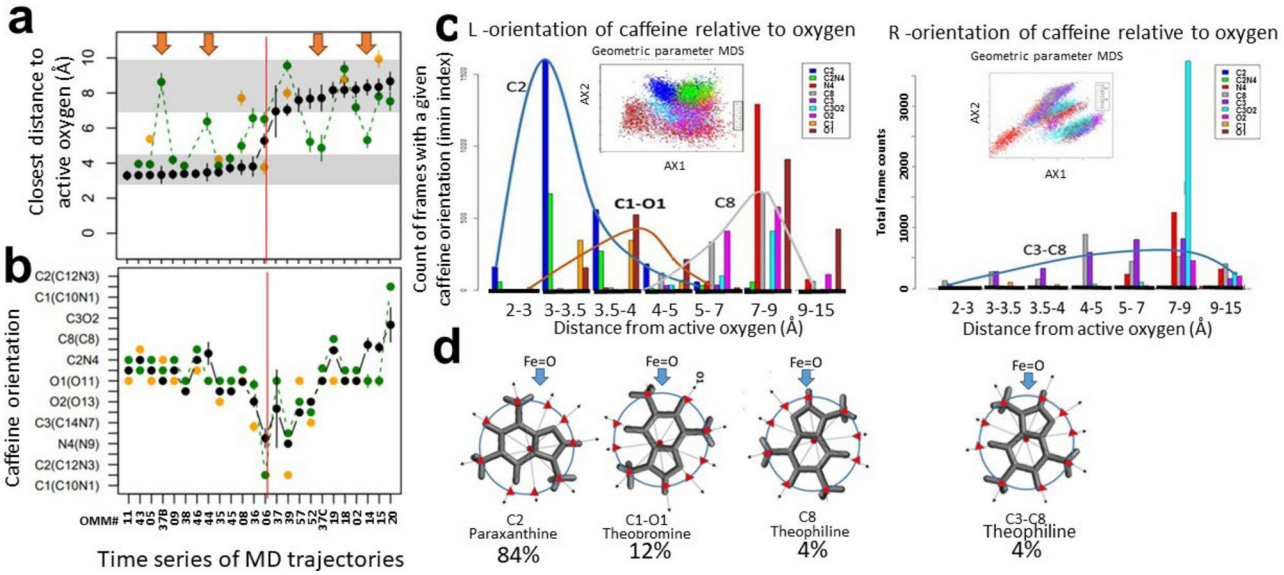

Fig. 4. Relationship between caffeine orientation and its distance to active oxygen. (**a**,**b**) Independent MD trajectories (X-axis) were converted into geometrical descriptors and their time-courses decomposed into sets of Gaussian distributions. The black, green, and yellow points represent the means of the major, secondary, and minor distributions, respectively, and their corresponding standard deviations. On the x-axis, trajectories were sorted by increasing values of the major Gaussian component of the Dmin time-series (black points on top left panel). Arrows point to trajectories involving substrate exchange between the producer and the effector sites (shaded areas). On panel (**b**) y-axis substrate orientations were repeated sequentially considering circularity. (**c**) Caffeine orientation distribution as function of its distance to active oxygen. MDs involving single occupancy at active site with L (left panel) or the R (right panel) substrate face configurations were considered. The caffeine atom closest to active oxygen was identified for each MD pose and corresponding counts plotted as function of the Dmin distance to active oxygen (x-axis). MDS maps of geometric descriptors corresponding to the same data are illustrated as inserts. (**d**) Respective caffeine orientation and expected corresponding metabolites are placed in front of the maximum of each distribution. Values are previously reported experimental proportions observed for each metabolite.

Several cases were observed in which a weakly populated Gaussian distribution (orange points in Fig. 4 and Fig. S5) pointed to transient geometries in which a caffeine molecule was transferring between the producer and the effector sites or reciprocally (see below). The impact of a potential stacking between caffeine and close phenylalanine residues was examined. Substrate orientation, distance to active oxygen (Imin, Dmin) and the angle formed by caffeine and the aromatic plane of Phe202 (226) appeared to be related (Fig. S6). This might result from the progressive release of the stacking-related constraints when the substrate was approaching producer site and acquiring its final orientation. To gain more insight, the Dmin distances were also mapped as a function of various combinations of geometric descriptors in Fig. S7. Such maps evidenced the alternate combinations of descriptors that were prone to enable the triggering of caffeine oxidation when the Dmin distance felt below 4 Å. Interestingly, the corresponding regions did not form a single domain but complex patterns, supporting that multiple combinations of geometric determinants can similarly trigger events. These maps supported in turn the capacity of defined geometric descriptors to direct machine learning processes integrating contributions of these multifactorial determinants.

Discontinuities in geometric parameters pointed to critical events for catalysis

Previous analysis illustrated that caffeine trajectories involved exchanges between fairly stable alternate geometries (long-life MD states). The presence of a sudden transition in values of A0, A2, D2 or stacking descriptors (Figs. S5, S6, S8), combined to observation of a substantial change in the D4 distance, was used to trigger us to focus on the corresponding time-series. Similar sequences of conformational events were evidenced, in which Phe-202 (226) functioned as a key (Fig. S9). This residue played a double role, acting as a door that isolates the two subsites or acting as a transfer helper, after a 90° rotation of its phenyl ring. As illustrated on Fig. S9, π–π stacking initiated with caffeine at the common border to producer and effector sites brought substrate to the effector site or reciprocally. Another type of remarkable event, leading to discontinuity in parameters, but without notable change in the D4 distance, could occur internally to the producer site (Fig. S10). This event corresponded to the rapid transition (in less than 1 ns) between alternate substrate orientations, resulting in similar positioning of a specific caffeine methyl group relative to active oxygen. This occurred particularly frequently for conformations that allow the formation of paraxanthine, the main caffeine metabolite (panel a of Fig. S10). As illustrated, the alternate configurations maintained optimal distances (~ 3 Å) between the carbon of the target methyl group and oxygen, enabling metabolism.

Caffeine rotamers: an exploration tool for the coupling between producer and effector subsites

The possibility of rotating the caffeine on its plane at both subsites without causing a major clash with protein side chains was exploited. The influence of the substrate orientation at one subsite on the dynamics of a sister molecule present at the other subsite was analyzed. Presence of a single substrate at the active site was used as a reference to analyze double occupancy conditions. A particularly stable double occupancy geometry previously observed during MD courses was used as a template to rationally design sets of rotamers at each subsite. Rotations by 30° steps were applied alternatively to caffeine at both subsites, keeping the centroid and the plane of the substrate unchanged (Fig. S11). Following these initializations, new MD were performed and analyzed. Two effects were observed (Fig. 5), one was the relaxation of the initial caffeine orientation in each subsite, and the other was the influence of the initial orientation in one subsite on the relaxation in the other subsite. In single occupancy conditions, partial relaxation of initial orientation settings occurred stepwise during short MD trajectories (~ 50 ns). Selection of a preferred orientation was more pronounced (mostly four preferred angles out of twelve rotamers) at the producer than at the effector site (Fig. 5 panels c, f). The latter site appeared to be more tolerant since initial orientation was maintained in nine cases out of twelve within the limits of Gaussian fluctuations. Globally, the presence of a second caffeine molecule affected preferred angles at each subsite (Fig. 5 panels a vs. c and e vs. f). However, the precise orientation set at the producer site did not specifically modify the relaxed orientation at the effector site (Fig. 5 panels e vs. b) while a weak but significant effect was observed in the mirror experiment (Fig. 5 panels a vs. d). Adding to orientation effects, we observed that during angular relaxations a limited but systematic shift of the caffeine plane occurred at the producer but not at the effector site (Fig. 5 panels g, h). We concluded that orientation constraints were much stronger at the producer site than at the effector site and that a weak but significant conformational coupling exists between the two subsites. The molecular basis of this coupling remained unclear, particularly since no significant contact between caffeine molecules was detected under these conditions.

Multidimensional scaling of geometrical parameters evidenced the presence of hidden rules that control substrate orientation

The orientation of caffeine in proximity of active oxygen (Imin or A2 angle) controls P450 regioselectivity and constitutes one of the six degrees of freedom that formally define the subsystem. However, molecular constraints related to protein side chains create relationships between descriptors. Constraints depended on the subsite

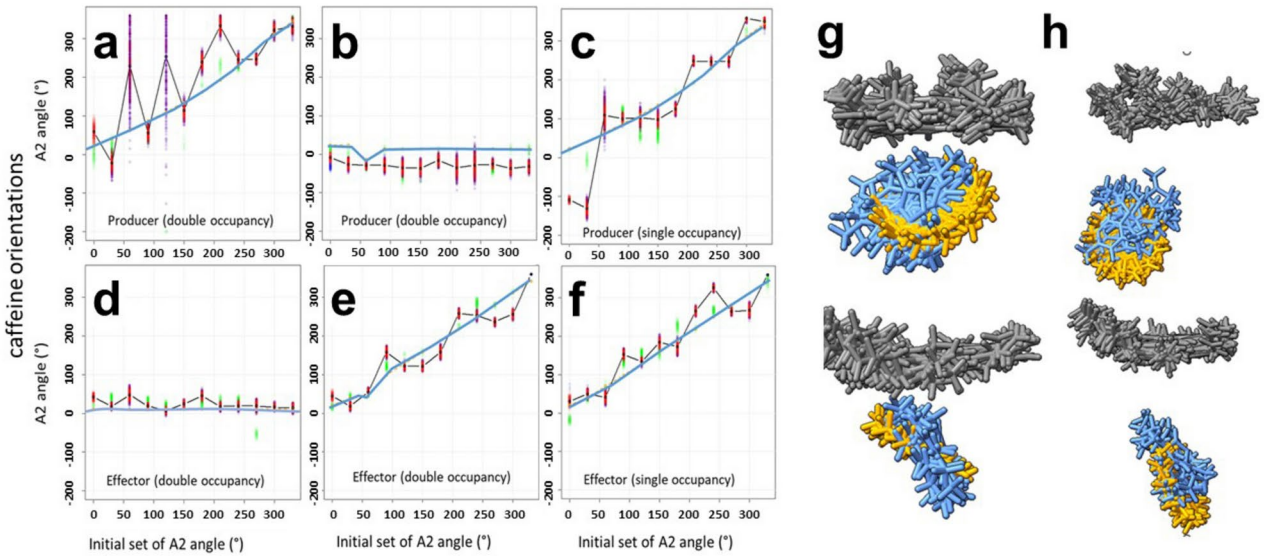

Fig. 5. Relaxation of initial caffeine orientations and coupling between substrate subsites in rationally designed rotamers. The x-axis of panels (**a**–**f**) indicates the initial orientations of caffeine (A2 angles) in the designed rotamers. Corresponding substrate orientations observed during MD relaxation are given on the y-axis for the substrate presents at the producer (**a**–**c**) or at the effector (**d**–**f**) site. When a second substrate was present (**a**,**b**,**d**,**e**), the angle relaxations were evaluated for the initially rotated substrate (**a**,**e**), and for the initially unchanged partners (**b**,**d**). Following MD initializations, the effective orientations observed during the first available MD poses can slightly differ from the designed orientations. The relations between these values are given by the blue lines. Orientations observed during relaxation were plotted as red dots for the major poses observed or as green dots for minor contributions. (**g**,**h**), caffeine geometries observed before (yellow) and following (blue) relaxations at producer (**g**) and effector (**h**) sites. All rotamers are overlaid and the corresponding heme poses are represented in gray color.

considered, the caffeine side facing the heme, and the eventual presence of a second caffeine molecule. To evaluate these, we reduced the number of descriptors by multidimensional scaling (MDS), optimally keeping their relative contributions to the variance of substrate orientations. MDS performed after descriptor centering and variance normalization allowed construction of bidimensional maps linking the reduced coordinates to the caffeine orientations (Fig. 6). A marked clustering of Imin indexes was observed in specific regions of these maps, particularly for Dmin distances that allowed catalytic actions (less than 3.5 Å). The comparison of maps supported that orientation mechanisms differed at the producer and effector sites (Fig. 6 panels c vs. d). Similarly, the maps differed for the R and L caffeine face configurations (panel a vs. e) and were modulated by the distance between the substrate and active oxygen (panels a vs. b and e vs. f). This last point was illustrated more precisely in Fig. 4 where the counts of caffeine orientations in the (R and L) face configurations were plotted as a function of the Dmin distance to active oxygen. The figure also illustrated that orientations in proximity of active oxygen seemed to have been progressively acquired during the substrate landing process and that orientation rules were specific for each substrate face. This point is of interest, as caffeine face flipping appeared impaired once the substrate was engaged in the active site channel. This suggested that factors influencing the selection of the substrate face outside the active site were finally affecting regiospecificity.

Design and scoring of a machine learning model that predicts caffeine orientation

The presence of complex but well-defined relations between geometric descriptors (A0, A1, A3, D2, D4) and substrate orientation (A2) enabled us to establish a machine learning model. This model was used in turn to evaluate the impact on regioselectivity of system perturbations. The rationale was to use the large sets of data available from MD trajectories and linking substrate orientation to its geometric context to train the model of Fig. S12. The effectiveness of training under different conditions was evaluated by scoring success rates by comparison of model predictions with corresponding taught MD data (see Methods). Comparison of learning efficiencies was used to evaluate how learning performed using data involving a given structural context applied to predictions in a different context (Fig. S13). This differential machine learning allowed us to define how factors like substrate face orientation or presence of a second caffeine molecule at active site impacted orientation mechanisms. Global scoring of predictions enabled some quantitative evaluation, while detailed analysis of correlation maps (Figs. 7 and 8) described how the different substrate orientations were affected. A set of control analysis was defined to evaluate potential bias or background of such an approach. In particular, a broken association between caffeine orientations and corresponding geometrical contexts was used during the learning phase to evaluate fortuitous matches. The ones resulting from the specific distribution of the learned orientations were first evaluated (Fig. S14, panel a1). Second, the orders of upstream (geometrical

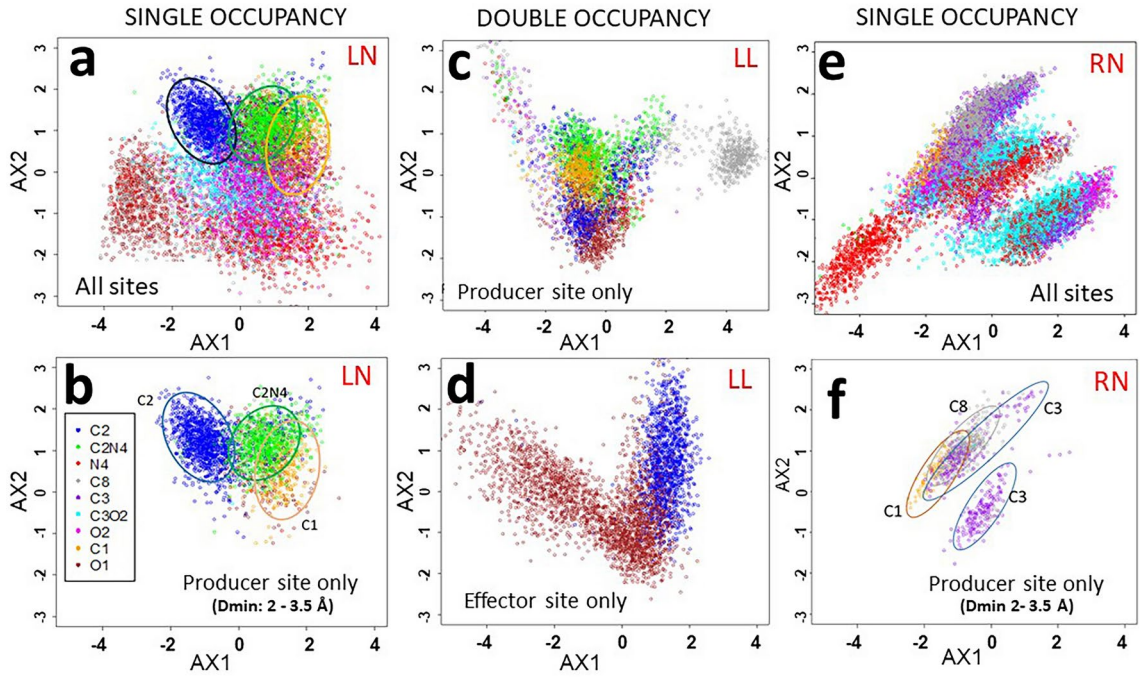

Fig. 6. Color-coded maps of caffeine orientations as a function of the MDS derived components of geometric descriptors. The set of geometric descriptors (A0, A1, A3, D2, D4) was reduced to two dimensions by MDS after variance normalization and centering. Reduced coordinates (ax1, ax2) of MD trajectories involving single (**a**,**b**,**e**,**f**) or double (**c**,**d**) site occupancies by caffeine with L (**a**–**d**) or R (**e**,**f**) face configurations were plotted. For each sampled MD pose (point), the caffeine atom closest to active oxygen was identified and color-coded as indicated in the inset. Corresponding distances were used for data filtering, in panels (**b**,**f**) (distance < 3.5 Å at the producer site). For panels (**c**,**d**), these distances were lower and greater than 5 Å, respectively. LN, LL, and RN stand for single occupancy L face, double occupancy L face, and single occupancy R face, respectively.

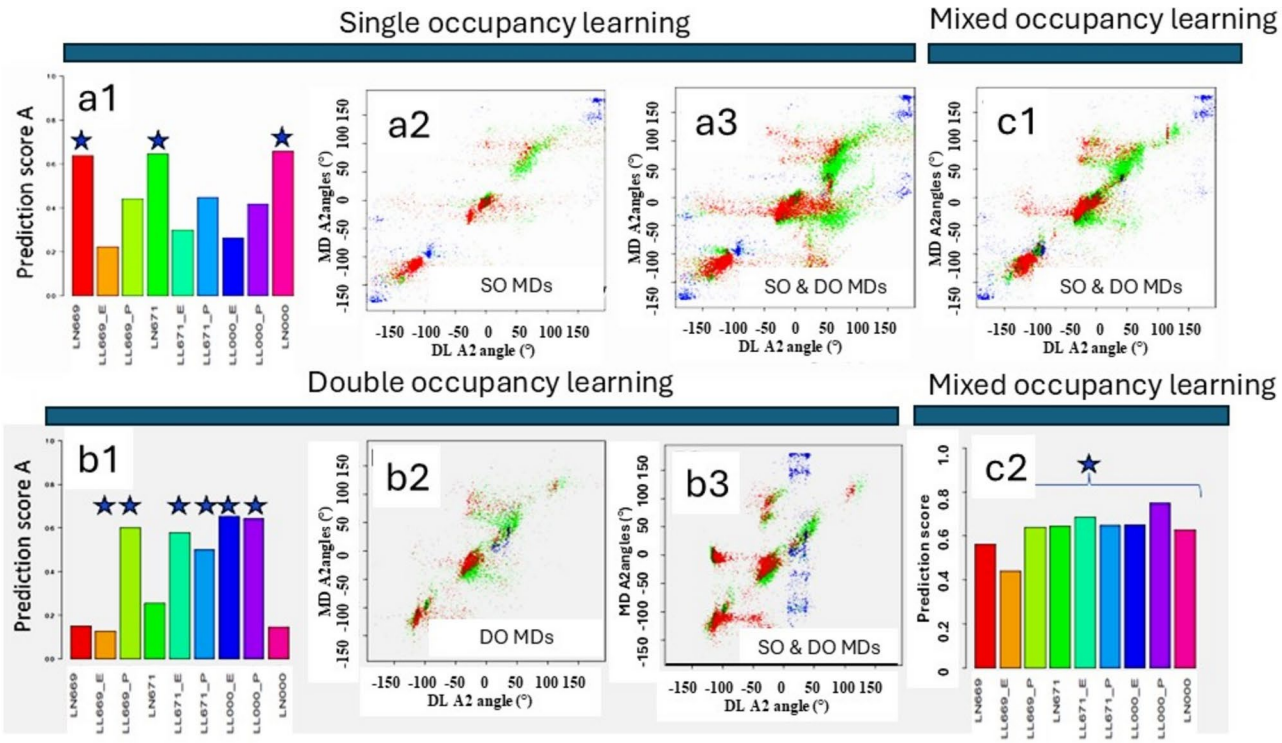

Fig. 7. Cross-validation of the predictive power of machine learning in single- and double-occupancy contexts. MD trajectories involving caffeine in the L face configuration in the single or alternatively double occupancy condition were used separately or in combination as learning or scoring data sets. MD defined values of A2 angles were plotted as a function of the machine learning predicted (DL) values (panels a2, a3, c1, b2, b3, c2). Color codes correspond to Dmin distances (red: < 3.5 Å, green: 3.5–5 Å, blue: > 5 Å). Learning was carried out using single (a1 to a3), double (b1 to b3) or mixed (c1, c2) occupancy data and the scoring was conducted using poses from single (a2), double (b2), all (a3, b3, c1) occupancy data. Success rates deduced from previous maps were presented as scores in panels a1, b1 and c2. Stars indicated the MD pools for which the learning and scoring sets involved the same level of occupancy (see supplementary tables for named sets on the x-axis). SO and DO stand for single and double occupancy evaluation sets, respectively. Bar colors corresponding to the same datasets were matched between panels a1, b1, c2.

context) or of downstream (orientation) data were randomized, breaking associations, and allowing evaluation of additional backgrounds (Fig. S14, panels b and c). Finally, the unmodified associations that enable learning were restored (panel d1) as a positive control. The results indicated that using a 15° tolerance for success in A2 angle predictions, fortuitous matches could not reach more than 20% of the tests in the worst possible case. This value was compared with scores in the range of 60–80% for the same datasets when unperturbed learning was applied. However, to further reduce background, the scoring method was improved by combining raw scores with a weighting factor derived from the pose-specific efficiency of learning (Methods and Fig. S15 panels a, e). This correction related on the sin/cos decomposition used to manipulate angle values that must have a constant module independent of predicted angles. This allowed us to sort correctly learned geometric contexts from incorrectly learn ones (S15 panels a, e) and to define corrections mostly canceling artifacts (panels d, h).

Contexts that affect substrate orientations at the active site are recognized by machine learning approaches

Comparative model scoring was used to analyze the rules driving substrate orientation. The rationale was to consider that identical or well-conserved rules would enable efficient predictions when using the learning set from one condition to predict caffeine orientations in another condition. In contrast, different rules would lead to poor or lack of correct predictions. This approach was used to better characterize the consequences of double caffeine occupancy and to evaluate how the caffeine side facing the heme can influence orientation mechanisms. For the first question, two batches of MD data were acquired, targeting rotamers at the producer and effector sites under double occupancy conditions. These parent data sets were used to generate two child datasets by alternatively deleting one or the other of the two substrates and restarting MD, which finally lead to 48 runs representing four sets of angle configurations and occupancy. Additionally, 24 (single occupancy) and 11 (double occupancy) MD runs without rotamers were considered as a complement. In all cases, the caffeine was set in the L face configuration. High and homogeneous success rates (around 60% except one case at 40%) were achieved for predictions when learning was performed using a mix of all data sets (Fig. S16, panels a-c). However, panels d-f indicated that these average success rates covered inhomogeneous contributions due to

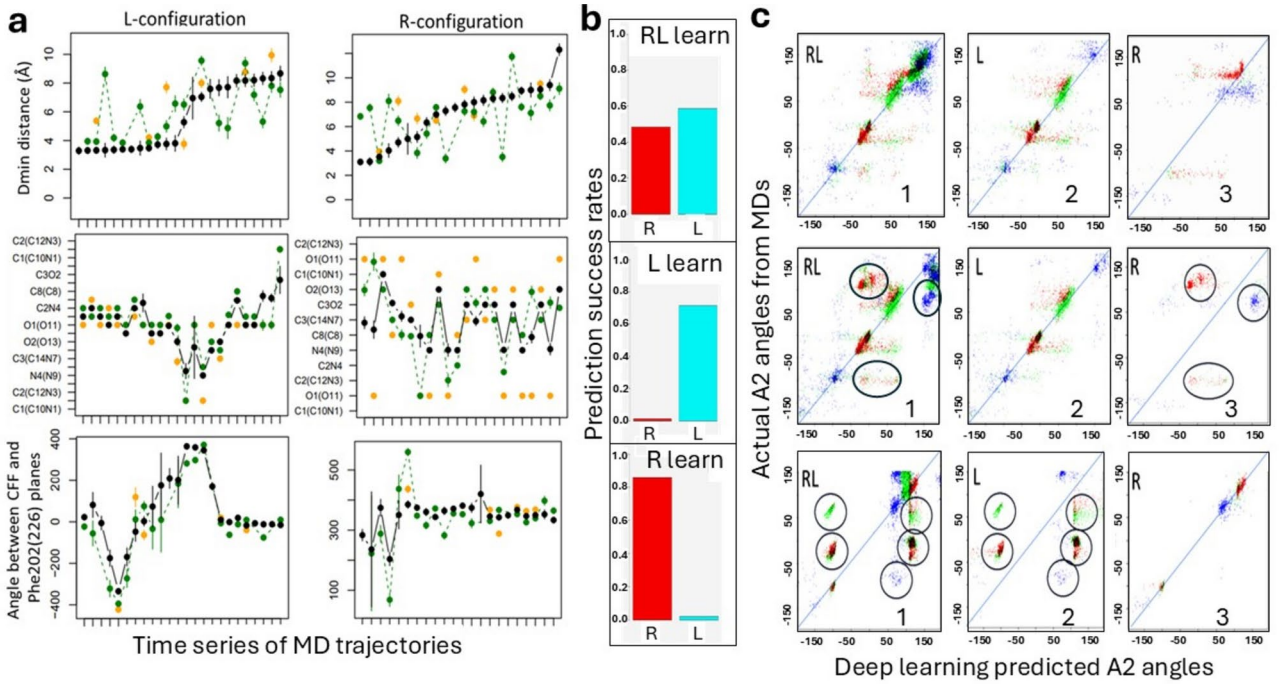

Fig. 8. Gaussian components of geometric descriptors and learning efficiencies for caffeine trajectories in the R and L face configurations. The trajectory of individual MD runs was converted into geometric descriptors and their values were decomposed into sets of Gaussian components. (**a**) Means and standard deviations for the Gaussian components (major: black, secondary: green, minor: yellow) of Dmin, Imin and the angle between caffeine and the aromatic plane of Phe-202 (226). The same order of trajectories (unnamed on x-axis) was used for all subpanels belonging to the same column of plots. Sorting was performed in order to sort the main components of Dmin values by increasing order in the top subpanels. (**b**) Success scores calculated from maps presented in panel **c** using data from the R (red) or L (blue) configuration for scoring and data from mix of R and L (top), L (middle) and R (bottom) configurations for learning. (**c**) MD defined A2 angles (y-axis) were plotted as a function of the corresponding machine learning predicted values (x-axis) in each subpanel. Color codes correspond to Dmin distances (red: < 3.5 Å, green: 3.5–5 Å, blue: > 5 Å). For the top, middle and bottom lines, learning was performed with data from a mix of R and L, only L or only R face configurations, respectively. For the left, middle, and right columns, scoring was performed using data from the mix of D and L, L only, or R only face configurations, respectively.

more efficient learning for frequently found A2 values. In the second step, learning was restricted to single or double occupancy conditions. Under these conditions, learning efficiencies appeared inhomogeneous and were significantly better when the same type (single or double) of occupancy was used for the learning and evaluation sets (Fig. 7). Noticeably, MD trajectories unrelated to rotamers resulted in intermediate scores, possibly because of a more limited representation in learning sets. A rather poor score was also observed in panel b1 for the LL669_E dataset, corresponding to rotamers at the effector site in double occupancy. However, better but still imperfect learning was observed in panel c2 for the same sample under mixed learning conditions. This observation could result from more limited orientation constraints at the effector site, leading to slower relaxation and sampling of unlearned transient states. Globally, orientation mechanisms that apply to single-occupancy cases poorly applied to double-occupancy conditions, supporting finding that the two subsites were interacting. This conclusion was reinforced by the data presented in Fig. S17, which presents analysis of the same data sets using an MDS approach. Maps illustrating substrate orientations in single- and double-occupancies were clearly divergent. Moreover, the decomposition by Dmin distances of the number of frames resulting in similar orientations evidenced that orientation constraints progressively changed as the substrate approached the producer subsite.

The same methodology was applied to test how the non-equivalence of the two caffeine faces may direct dissimilar regioselectivity. Sets of MD trajectories involving caffeine in the R or L face configurations were acquired, converted into geometrical descriptors, and used for machine learning. The ability of models taught in one configuration to predict caffeine orientations in the other configuration was evaluated using scoring as previously described (Fig. 8, panel c). The results evidenced that efficient predictions were obtained when learning and scoring were performed using MD data obtained with the same face configuration but that poor or irrelevant results were achieved in the contrary. Interestingly, learning using a mix of face configurations (Fig. 8, panel c, top subpanels) allowed the model to reasonably predict substrate orientation regardless of the face considered. This confirmed the ability of the model to learn rules associated with the two configurations simultaneously. In contrast, these rules differed for models separately taught with trajectories involving caffeine

in alternate face configurations (Fig. 8, panel c, middle, and bottom subpanels). Thus, the mechanisms that define the substrate orientation relative to active oxygen clearly differed for the two faces of caffeine. These changes did not result in a simple mirroring of the preferred orientations due to the absence of symmetry axis in the caffeine molecule (Fig. 8, most right maps of panel c). Consistently, the relationship between Imin and Dmin values observed with caffeine in the L face orientation was no longer conserved in the R configuration (Fig. 8, panel a, b). The same was applied when the relationship of Imin value with the angle formed by Phe-202 (226) and caffeine planes was considered (Fig. 8, bottom of panel a). Thus, the regioselectivity resulting from the combination of the Imin and Dmin values is expected to differ markedly depending on the configuration of the caffeine face in the active site channel.

Discussion

MD coupled to quantum chemistry is a unique tool to explore active site reactions and to capture elementary events that control substrate binding, orientation, and finally product formation. However, the approach suffers from the difficulty of extracting mechanisms underlying complex trajectories and exhaustively sampling the broad range of possible situations. Conversion of time-series of coordinates into sets of geometrical descriptors was successfully performed by different generalist methods[24] including time series of 2D-tensors[21], equivariant geometry-enhanced graph neural network[20] or geometry-enhanced molecular representation[25]. Despite their power, such approaches were not designed to encode specific system features. Thus, they have learning, predictive, and descriptive qualities, but mostly lack power to evidence underlying mechanistic rules. In contrast, our P450 dedicated approach was not generalist while offering ways for transposition to other substrates or enzymes in which a tightly bound cofactor plays a leading role in catalytic actions.

A second level of encoding was applied following observation that geometric descriptors derived from substrate trajectories involved frequently stepwise transitions between states characterized by Gaussian fluctuations around long-life geometries. Assuming the frequent occurrence of such states and for similar computational efforts, multiple independent and relatively short MD runs allowed us to efficiently sample substrate trajectories in a potentially better way than a more limited number of longer dynamics. The network of relationships evidenced between the resulting geometric descriptors proved to be sufficient to direct efficient machine learning. Resulting models were used into an unusual manner as a tool to analyze how orientation rules learned in a structural context can or not be applied to another structural context. This way proven to be efficient to evaluate how orientation mechanisms depended on the substrate face and subsite interactions. Although exemplified here for specific cases, this approach could be generalized to characterize mechanistic consequences of any structural modifier, including factors acting outside the active site, like distant mutations, lipid environment, or access channel structure.

Changes in the P450 1A2 metabolite pattern at high caffeine concentration[5] as well as previous in silico docking and MD studies[4,11] supported the presence of multiple binding sites[26]. However, their functional roles remained unclear as caffeine saturation curve at low concentrations mostly exhibits a Michaelian aspect. This could appear contradictory with cooperative effects classically expected when two interacting binding sites are present close to active site[27]. However, in our models, a more complex situation occurred involving a multistep substrate orientation mechanism as summarized in Fig. S18. The substrate presence at the producer site was obligatory sequential to a previous occupancy at the effector site, and substrate transfer between the two sites was gated by Phe-202 (226) and more marginally by Phe-236 (260). The same likely applies in the opposite direction for product release. At the effector site, the caffeine centroid was typically distant by ~ 12 Å from active oxygen, preventing formation of metabolites, in contrast to the situation at the producer site, where this distance was in the range of 3 Å. Interestingly, we illustrated that alternate and long-life caffeine geometries relative to heme can result in identical positioning of the same methyl group relative to active oxygen. This implied that alternate geometries could lead with potentially similar efficiencies to the production of the same metabolite. Globally, classical allosteric models can hardly be applied to such a complex mechanism that makes effects of site interactions on saturation curves hard to predict.

Substrate exchanges between the producer and effector sites were found to occur in the 10–50 ns range, thus involving a limited activation energy and rather equilibrated subsite occupancies considering predicted affinities. However, substrate present at producer site appeared more prone to escape toward effector site than the contrary. Rotamer relaxation analysis taught us that orientation bringing methyl C2 (12) in the vicinity of active oxygen, thus favoring the formation of the main metabolite, appeared particularly stable even when effector site was unoccupied.

Regarding the effector site, Phe-202 (226) and, more marginally, Phe-236 (260) played key roles. Substrate stacking with Phe-202 was the main driving force stabilizing substrate at this subsite. This location was particularly prone to permit substrate reorientation by rotation on its own plane without disrupting stacking. Consistently, rotamer analysis evidenced only weak preferences during relaxation of preset orientations. The role of the stacking with Phe-236 (260) was in contrast unclear, considering its very variable extent for similar substrate positioning. Events of substrate transfer between the producer and the effector sites were relatively infrequently observed, but when present occurred abruptly in a time range of few ns and were gated by Phe-202 (226). Consistent with the narrow geometry of active site channel, caffeine escape from the producer site was never observed when the effector site was occupied. In the absence of substrate at the effector site, observed residence time of caffeine at the producer site was highly dependent on its initial orientation. In contrast, presence of a second molecule locked the substrate at the producer site, giving it time to eventually relax toward a more stable orientation, which in turn could affect regioselectivity. This lock-and-relax mechanism made the caffeine orientation under the thermodynamic control of its relaxation at producer site when the effector site was occupied. In the contrary, the orientation dependence of the kinetic of substrate escape may limit the possibilities of relaxation making regioselectivity more on kinetic control. Interestingly, Phe-202 (226) which stabilizes the

effector site played an alternate role in initiating or maintaining a partial substrate stacking during transfer, likely reducing the energetic barrier. External trigger events controlling this mechanism were not evidenced but may in part be related to the more stochastic fluctuation of Phe-236 (260).

Differential machine learning and MDS analysis evidenced that caffeine orientation rules clearly differed for the two faces of substrate and that only one orientation could preferentially led to formation of the major metabolite. This difference was consistent with the absence of symmetry axis on the caffeine molecule. Face selection must occur before substrate entry into the flat active site channel, likely at level of more distant tunnels or at the protein-membrane interface. Such mechanism is particularly hard to analyze considering the multiplicity of potential channels and that caffeine dipolar moment is mostly oriented on the axis of substrate plane. Surprisingly, none of the previous reports pointed out or even took this problem into account. Globally, the described geometric approach can be transposed to other P450s or cofactor containing enzymes. In the case of P450s, it can be generalized to address the important question of polymorphism or metabolic interference when several molecules can bind the enzyme.

Methods

Caffeine atoms and amino-acid numbering

Due to software constraints imposed by caffeine parametrization followed by automated data processing of large data sets, the atom numbering in this study differed from the standard caffeine atom numbering as defined in Fig. 2. For similar reasons and the use of amino acid sequences truncated from their signal peptides, the numbering of residues differed from the full-length protein ones. Such a numbering scheme was used in supporting scripts and raw data files. However, to facilitate reading, numberings relative to the full-length sequence were also given in brackets throughout the text.

Cofactors and substrate parameters

The complex between human P450 1A2 and the FMN domain of the reductase was modelled as previously described[25]. Flavin geometry optimization involved the GAMESS software (https://www.msg.chem.iastate.edu/gamess/) and charge calculation the #B3LYP/6-311G level of theory with cofactor environment customized using solvent parameters that simulate an average protein environment (Eps = 9.2, n = 1.51, Vmol = 55, density = 0.033). Amber tools (https://ambermd.org/AmberTools.php) were used to convert quantum mechanics calculation outputs to RESP charges and mol2 files according to archive.ambermd.org/200812/att-0116/resp_fit.htm. Heme cofactor parameters for the ferryl-oxo complex were evaluated using the same tools and B3LYP/6-31G* (charge -2, multiplicity 4) theory level. Conversion to Amber parameters was performed using the MCBP.py Python (bugs corrected) routine according to Amber tools tutorial[28] for heme group building (https://ambermd.org/tutorials/advanced/tutorial20/mcpbpy_heme.php). The caffeine parameters were established using the Amber two-step procedure for ligands, B3LYP/6–311 + + G** level of theory and the same tuned solvent environment. Amber tools were used for the final setting of MD. UCSF Chimera, ChimeraX, VMD and homemade python scripts using the UCSF chimera python library were used for additional format conversions and analysis.

Molecular dynamics simulations

MD runs were set in an explicit 0.1 M NaCl pH 7.0 solvent and performed using Ambertools 22 and OpenMM software (https://openmm.org/) with CUDA implementation. Non-bounded PME non-bounded methods (cutoff 1.0 nm), hydrogen mass repartitioning (1.5 amu), step size of 5 fs, constant pressure (1 atm) and temperature (300°K) were used. Simulation lengths were adjusted according to purpose without generally exceeding 70 ns.

Data analysis

Data analysis involves MD description files given in tables S4 and S5 and was performed using R-scripts and complementary files deposited in the GitHub repository under the publicly available 'Caff-1A2-Rscripts' entry. Raw molecular dynamics data, pre-processed files and associated figures were deposited as zip compressed folders at https://doi.org/10.57745/AEB7SV under reference P450MolDyn-DP2024.The local coordinate system used was based on the average plane formed by the four iron porphyrin nitrogen ligands and their mean projections on the normal vector to this plane (Fig. 3 and Fig. S3). The signs of calculated angles were relative to the N1-N4 rotation in the porphyrin ring plane and to the C1-C3 (C10-C14) rotation in the substrate plane. However, reconciliation of these two local references was related to the placement of the iron/active oxygen projection onto the caffeine plane relative to the intersect line of the heme and caffeine planes. When crossing this line, angle A3 becomes shifted by 180°, introducing irrelevant discontinuities. The trick was to encode such a crossing in the sign of the D2 distance while maintaining the continuity on the A3 angle values. An additional correction was used to consider the side of the caffeine molecule that faces porphyrin into the active site channel. This correction maintains angles A0, A1, A3 as well as distances D2 and D4 unchanged when caffeine was reflected in its plane and the centroid of its methyl groups kept unchanged. For all dynamics, conversions of MD coordinates to geometrical descriptions were secured in the script by the reverse conversion to classical atom coordinates and comparison.

Statistical analysis of geometric parameters was automated, including when multimodal behaviors were present. For a given parameter, each time series was clustered into sets of Gaussian distributions. The validity of the resulting decomposition was evaluated using statistical criteria (adjustment to the law, fit coverage, and separation of modes). Minor subpopulations (less than 5%) were excluded during the procedure. The routines used are also part of the R scripts released.

Multidimensional scaling
Time-series of atom coordinates (excluding the first ns to allow initial relaxation) were converted into geometrical parameters. Descriptors (excluding substrate orientation) were centered and normalized on their variances before being reduced to two dimensions of the parameter sets using an MDS algorithm (see R code). The substrate orientation (A2) was separately color-encoded into points on the 2D maps.

Machine learning
Data used for machine learning analysis followed a similar setting procedure up to the variance normalization step before being split into a learning sub-population (70%) and a test sub-population (30%). A KERAS model including five layers with 9, 128, 64, 12 and 2 neurons (RELU activations) supplemented by a dropout layer (rate = 0.3) between layers 2 and 3 of neurons was designed by trial and error. This model did not rely on the time sequence of MD data that was shuffled in a random order prior to use. To avoid perturbation by the cyclic nature of angle definition, all angles were converted to pairs of sinus and cosine values, increasing the number of entry parameters from 5 to 8 ($2 \times 3 + 2$). Similarly, caffeine orientation angles (A2) were converted into trigonometric pairs of values. The learning steps were always performed using these sin/cos decompositions, and the reverse conversion to angles was performed following the prediction from machine learning.

Scoring of machine learning predictions
Raw scoring was performed based on the success rates (predicted orientation within a 15° range from the MD-defined A2 angle). The modules of predicted cos/sin decompositions of A2 angles were also considered. Such modules were expected to be unit irrespectively of A2 values in case of correct predictions while remaining undefined in case of failed learning. Such failure was simulated by teaching the model using randomized associations of context and target values. In such condition, the sin/cos modules of predicted A2 values formed a Gaussian distribution centered around 0.35. Consequently, raw success scores were weighted by a function of predicted sin/cos modules designed to cancel contribution of poorly learned poses (see scripts for details). As a result, corrected scores ranged between 0 and 1 for absent and efficient learning, respectively.

Caffeine rotamers constructions
Alternate orientation settings (stated as rotamers in the results) of the caffeine molecule were built from the same template structure through R scripts coupled to Python routines allowing automatic regeneration of Amber parameters used in dynamics. Caffeine at the producer or effector site was rotated by a clockwise increment of 30° around the center of the circle circumscribing its methyl groups, keeping its plane unchanged. Initial settings (set as red dots in the figures) were first relaxed for 1 ns before running longer dynamics and analyzing. In case of double occupancy, a single caffeine was rotated at a time and the second kept unmodified. The unique template pose selected to generate all of the rotamers corresponded to a particularly metastable geometry of the two substrates that was identified among the runs performed in double occupancy conditions. Commentated R-scripts organized in functional modules were used for all calculations and are available as GITHUB deposits. The code can be adapted to address comparable questions about other enzymes involving catalytic cofactors or different substrates.

Angle labeling on figures
To improve readability by maintaining as possible trace continuities in figures involving angles, an offset by multiple of 360° was applied to the raw angle values. However, this correction was data and figure specific and must be kept in mind as possibly introducing non-relevant visual discrepancies when comparing values in between figures.

Data availability
R-scripts used for data analysis including machine learning are publicly available in the GitHub repository under the 'Caff-1A2-Rscripts' entry. The description files required by the scripts (except for the raw molecular dynamics data) were deposited at the same place. Molecular dynamics data in a format compatible with scripts, preprocessed files and sets of figures associated with each molecular dynamic trajectory were deposited as two zip compressed folders on URL: https://doi.org/10.57745/AEB7SV under reference P450MolDyn-DP2024 (INRAE entrepot).

Received: 3 October 2024; Accepted: 18 February 2025
Published online: 03 March 2025

References
1. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143 (2018).
2. Meng, S. et al. The molecular basis and enzyme engineering strategies for improvement of coupling efficiency in cytochrome P450s. *Biotechnol. Adv.* **61**, 108051 (2022).
3. Mukherjee, G., Nandekar, P. P. & Wade, R. C. An electron transfer competent structural ensemble of membrane-bound cytochrome P450 1A1 and cytochrome P450 oxidoreductase. *Commun. Biol.* **4**, 55 (2021).
4. Sun, M., Lyu, L. & Zheng, Q. Active binding modes of caffeine with cytochrome P450 1A2 determine its metabolite profiles. *Chem. Res. Toxicol.* **36**, 1313–1320 (2023).
5. Regal, K. A. & Nelson, S. D. Orientation of caffeine within the active site of human cytochrome P450 1A2 based on NMR longitudinal (T1) relaxation measurements. *Arch. Biochem. Biophys.* **384**, 47–58 (2000).
6. Cameron, M. D. et al. Cooperative binding of acetaminophen and caffeine within the P450 3A4 active site. *Chem. Res. Toxicol.* **20**, 1434–1441 (2007).

7. Sohl, C. D. et al. Cooperativity in oxidation reactions catalyzed by Cytochrome P450 1A2 highly cooperative pyrenehydroxylation and multiphasic kinetic and ligand binding. *J. Biol. Chem.* **283**(11), 7293–7308 (2008).
8. Harrelson, J. P., Atkins, W. M. & Nelson, S. D. Multiple-ligand binding in CYP2A6: probing mechanisms of cytochrome P450 cooperativity by assessing substrate dynamics. *Biochemistry* **47**, 2978–2988 (2008).
9. Strohmaier, S. J., De Voss, J. J., Jurva, U., Andersson, S. & Gillam, E. M. J. Oxygen surrogate systems for supporting human drug-metabolizing cytochrome P450 enzymes. *Drug Metab. Dispos.* **48**, 432–437 (2020).
10. Urban, P., Lautier, T., Pompon, D. & Truan, G. Ligand access channels in cytochrome P450 Enzymes: A review. *Int. J. Mol. Sci.* **19**, 1617 (2018).
11. Jandova, Z., Gill, S. C., Lim, N. M., Mobley, D. L. & Oostenbrink, C. Binding modes and metabolism of caffeine. *Chem. Res. Toxicol.* **32**, 1374–1383 (2019).
12. Zhang, T., Liu, L. A., Lewis, D. F. V. & Wei, D.-Q. Long-range effects of a peripheral mutation on the enzymatic activity of cytochrome P450 1A2. *J. Chem. Inf. Model.* **51**, 1336–1346 (2011).
13. Mokkawes, T. & de Visser, S. P. Caffeine biodegradation by cytochrome P450 1A2. What determines the product distributions?. *Chemistry* **29**, e202203875 (2023).
14. Prašnikar, E., Ljubič, M., Perdih, A. & Borišek, J. Machine learning heralding a new development phase in molecular dynamics simulations. In *Artificial Intelligence Review* (Springer, 2024).
15. He, Y. et al. Using deep learning and molecular dynamics simulations to unravel the regulation mechanism of peptides as noncompetitive inhibitor of xanthine oxidase. *Sci. Rep.* **14**, 174 (2024).
16. Sun, Y., Jiao, Y., Shi, C. & Zhang, Y. Deep learning-based molecular dynamics simulation for structure-based drug design against SARS-CoV-2. *Comput. Struct. Biotechnol. J.* **20**, 5014–5027 (2022).
17. Chang, J., Fan, X. & Tian, B. DeepP450: Predicting human P450 activities of small molecules by integrating pretrained protein language model and molecular representation. *J. Chem. Inf. Model.* **64**, 3149–3160 (2024).
18. Tsai, S. T., Kuo, E. J. & Tiwary, P. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nat. Commun.* **11**, 1–11 (2020).
19. Yu, L., He, X., Fang, X., Liu, L. & Liu, J. Deep learning with geometry-enhanced molecular representation for augmentation of large-scale docking-based virtual screening. *J. Chem. Inf. Model.* **63**, 6501–6514 (2023).
20. Wang, Y. et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat. Commun.* **15**, 313 (2024).
21. Berishvili, V. P. et al. Time-domain analysis of molecular dynamics trajectories using deep neural networks: Application to activity ranking of tankyrase inhibitors. *J. Chem. Inf. Model.* **59**, 3519–3532 (2019).
22. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
23. Esteves, F. et al. Single mutations in cytochrome P450 oxidoreductase can alter the specificity of human cytochrome P450 1A2-mediated caffeine metabolism. *Biomolecules* **13**, 1083 (2023).
24. Dvořák, J., Mañák, M. & Váša, L. Predictive compression of molecular dynamics trajectories. *J. Mol. Graph. Model.* **96**, 107531 (2020).
25. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
26. Korzekwa, K. R. et al. Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry* **37**, 4137–4147 (1998).
27. Davydov, D. R. & Halpert, J. R. Allosteric P450 mechanisms: multiple binding sites, multiple conformers or both?. *Expert Opin. Drug Metab. Toxicol.* **4**, 1523–1535 (2008).
28. Li, P. & Merz, K. M. MCPB.py: A python based metal center parameter builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).

Author contributions

D.P. designed the whole work strategy, molecular dynamic models, geometric encoding rules, and machine learning approaches, wrote R-scripts, performed calculations, statistical analysis, and drafted the initial paper draft. P.U.and L.G. made important suggestions to improve data analysis and paper draft and contribute to define and to analyze available state of the art. The last two authors had similar contributions.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91155-0.

Correspondence and requests for materials should be addressed to D.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.