It's Not Always Crystal Clear: In-Pocket Analysis for Understanding Binding Site Microenvironment Properties

Filipe Menezes,[1,2,*] Tony Fröhlich,[1,2] Lucy C. Schofield,[3] Julian Cremer,[4] Jon Agirre,[3] Robbie P. Joosten,[5,6] Garib Murshudov,[7] Oliver Plettenburg,[8,9,10,11] Michael Sattler,[1,2] Grzegorz M. Popowicz[1,2,12*]

[1]Molecular Targets and Therapeutics Center, Institute of Structural Biology, Helmholtz Munich, Neuherberg, Germany.

[2]TUM School of Natural Sciences, Department of Bioscience, Bayerisches NMR Zentrum, Technical University of Munich, Garching, Germany.

[3]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England.

[4]Machine Learning & Computational Sciences, Pfizer Worldwide R&D, Berlin, Germany

[5]Department of Biochemistry, Netherlands Cancer Institute, Amsterdam, the Netherlands.

[6]Oncode Institute, the Netherlands.

[7]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge, CB2 0QH, United Kingdom.

[8]Molecular Targets and Therapeutics Center, Institute of Medicinal Chemistry, Helmholtz Munich, Neuherberg, Germany.

[9]Institute of Organic Chemistry, Center of Biomolecular Drug Research (BMWZ) and Laboratory for Quantum and Nano Engineering (LNQE), Leibniz Universität Hannover, Hannover, Germany.

[10]Institute for Lung Health (ILH), Giessen, Germany.

[11]German Center for Diabetes Research (DZD), Neuherberg, Germany.

[12]Khumbu.Ai GmbH

ABSTRACT

Structural biology has enabled key advances in biochemistry, drug discovery, and AI-driven three-dimensional structure modelling. Despite its centrality, no general and reliable computational tools exist to chemically validate structural models, especially when small-molecule ligands are involved. This challenge is exacerbated when experimental resolution is insufficient to define interaction chemistry, leaving critical interpretations ambiguous. We introduce In-Pocket, a quantum mechanical optimisation method biased by experimental data that resolves the chemical nature of protein-ligand binding interfaces. In-Pocket improves model quality, corrects chemically inconsistent structures, and identifies the energetically most favourable conformers within biomacromolecular complexes. Further, it generalises across a wide range of systems, including those containing transition metals. Benchmarking demonstrates that In-Pocket enhances the reliability of legacy and newly solved structures. With automated and simple-to-use routine chemical validation, In-Pocket stands to improve the integrity of structural databases and impact diverse domains: from molecular biology to rational drug design and application of generative AI.

INTRODUCTION

The Protein Data Bank (PDB)[1] is the reference repository for biological macromolecular structures. This foundational community resource comprises over 235,000 entries and has been instrumental in enabling the development of transformative machine learning (ML) approaches such as AlphaFold[2] or RoseTTAFold[3]. The majority of PDB structures were determined using macromolecular X-ray crystallography (MX),[4] with additional contributions from nuclear magnetic resonance (NMR) and the growing single particle cryo-electron microscopy (cryo-EM).[5]

All current methods are fundamentally constrained in resolution. Other than infrequent sub-Ångström resolution cases, typical MX and cryo-EM limitations include: i) reduced precision in resolving bond lengths and angles in the heteroatom scaffolds;[6,7] ii) widespread absence of experimental evidence for hydrogen atoms; and iii) frequent ambiguity in the interpretation of density maps. Further, canonical representations of molecular scaffolds fall into the traditional and simplistic "ball-and-stick" category, which misses the true molecular electronic structure. Improper ligand structural assignments, where inaccurate bond assignments dictate the geometry and strength of binding interactions, propagate significant downstream errors. Ultimately, this compromises the reliability and usability of structural models, especially for binding strength prediction, which is essential for drug discovery, and in training novel generations of ML models. Frustratingly, chemically informed computational validation of biomolecular models has remained a critical unmet need in structural biology for years.

To address this, we introduce In-Pocket Analysis (IPA), an algorithm for evaluating the local chemical environment of a compound or residue within a macromolecular structure (**Figure 1a-d**). While often focused on small-molecule ligands, IPA is equally applicable to amino acids and their modifications, nucleic acids, and metal-containing cofactors. The protocol performs experimentally biased quantum-mechanical (QM) optimisation of ligand-residue interactions across multiple protonation and tautomeric states–termed chemical states. Through validation against deposited experimental data, we show that IPA is a robust, quantitative strategy for improving model accuracy and chemical interpretation by identifying the most chemically plausible states within a given structural context.

RESULTS

To correct the structural model, IPA assesses ligand-environment-element interaction energies (E_{int}), the ligand deformation energy (E_{def})–the receptor-induced energetic cost required for a free ligand in solution to adopt its bound conformation–and the ligand distortion energy (E_{dist}), which arises from the accumulation of experimentally driven structural inaccuracies, *i.e.*, those caused by insufficient resolution. These quantities are automatically evaluated for several ligand chemical states and integrated into a binding energy ($E_{bind}=E_{int}+E_{def}$), facilitating more accurate structure-based modelling. Supplement **S1** provides a theoretical rationale for E_{dist} and how it is used in downstream analysis.

Protein-ligand complex databases are contaminated with structural distortion.

The binding of a ligand to a protein entails dovetailing interactions that reshape both partners, often accompanied by structural distortion energy. However, modelling inaccuracies caused by the finite resolution of experimental structural data lead to artefactual, overly large distortion energies. Here, we used IPA to evaluate statistical metrics of distortion and deformation across a protein-ligand structure database (**Figure 1e**).[8,9] Our analysis reveals a narrow distribution of ligand deformation energies, peaking around 5 kcal/mol and quickly decaying. Conversely, the distortion distribution is broader and peaks between 10-20 kcal/mol, with a long tail extending well beyond 1,000 kcal/mol. These findings show

that distortion is 2-4 times greater than the deformation energy required for binding, highlighting substantial inaccuracies in structural models. Such distortions translate into misassigned atomic hybridisation and protonation states when neglecting PDB-derived topologies.[9]

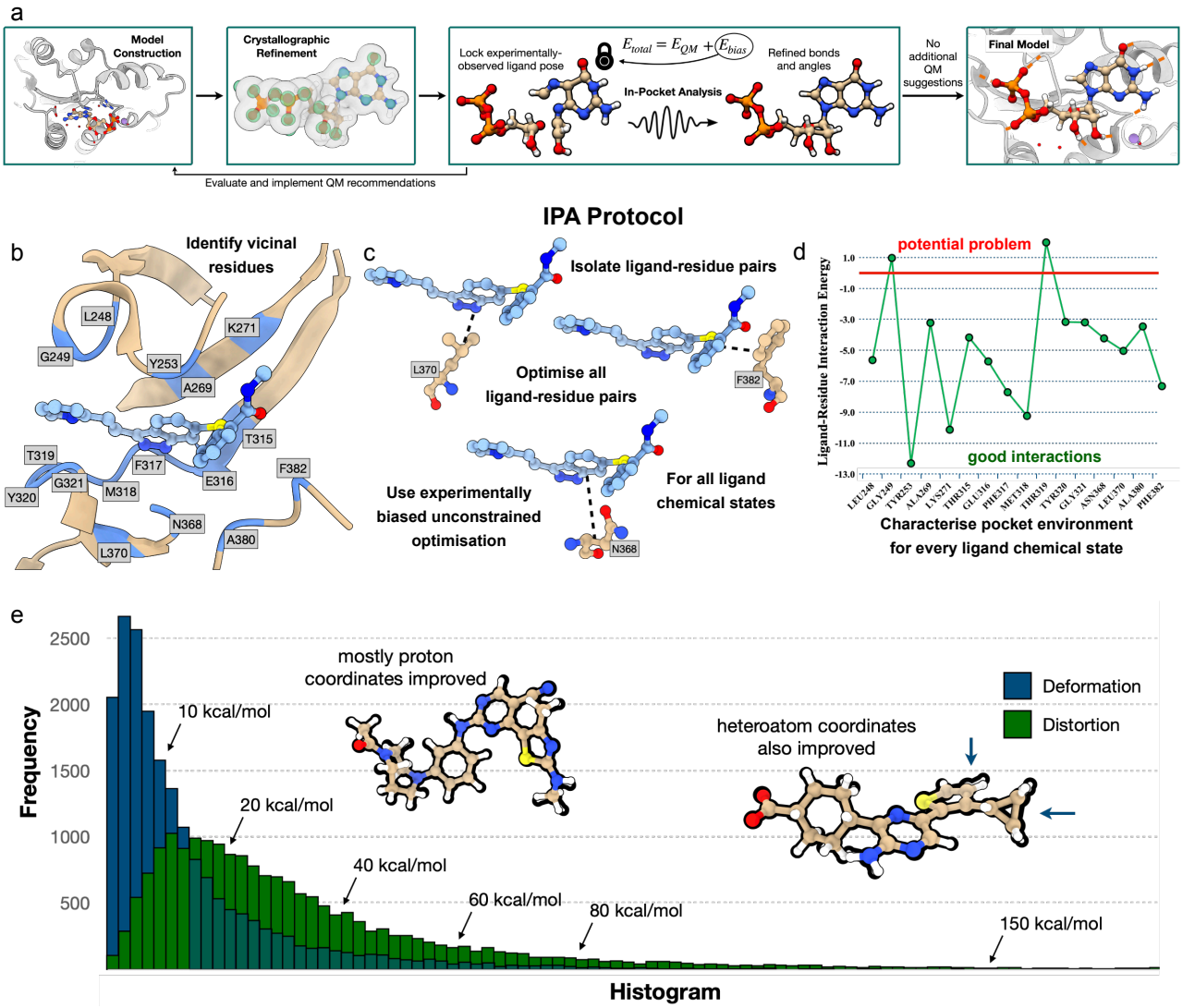


Figure 1: IPA for detailed chemical examination of biological macrocomplexes. a-d) The IPA protocol. a) Modified crystallographic model-solving pipeline with embedded IPA for chemically refined structural models. Given a biological structural model, b) IPA identifies all residues within the vicinity of the active residue, typically a small molecule ligand. c) IPA then uses a series of experimentally biased QM optimisations of each molecular pair to d) calculate experimentally relevant interaction energies. When negative, these energies indicate good, attractive interaction pairs. On the other hand, positive interaction energies indicate possible pathologies in the structural model. e) IPA is used to analyse experimental model distortion and the ligand deformation energy of 20,000 ligands available in binding databases.[8,9] Two different histograms are obtained: a distribution peaked at about 5.0 kcal/mol, with a quick decay, is obtained for IPA-derived ligand deformation energies. A broader distribution peaking at 10-20 kcal/mol, slower decay and an extremely long tail is obtained for the experimental distortion energies.

The origins of structural distortion are multifold. At the lower end, inaccuracies in proton coordinates and minor angular distortions are recorded. We observed short contacts between hydrogen atoms, substantial angular distortions, or weak bond-length distortions for medium-range distortion energies. For instance, all nitro-containing compounds we analysed contained N-O bonds deviating by 0.1-0.2 Å from the reference value of 1.22 Å.[10,11] We observed molecules with severe

bond and angular distortion in the high-distortion domain. Typical cases include bonds that are too short or overly stretched, often in the same structure. Such examples are analysed later in the text (see **Figure 5**).

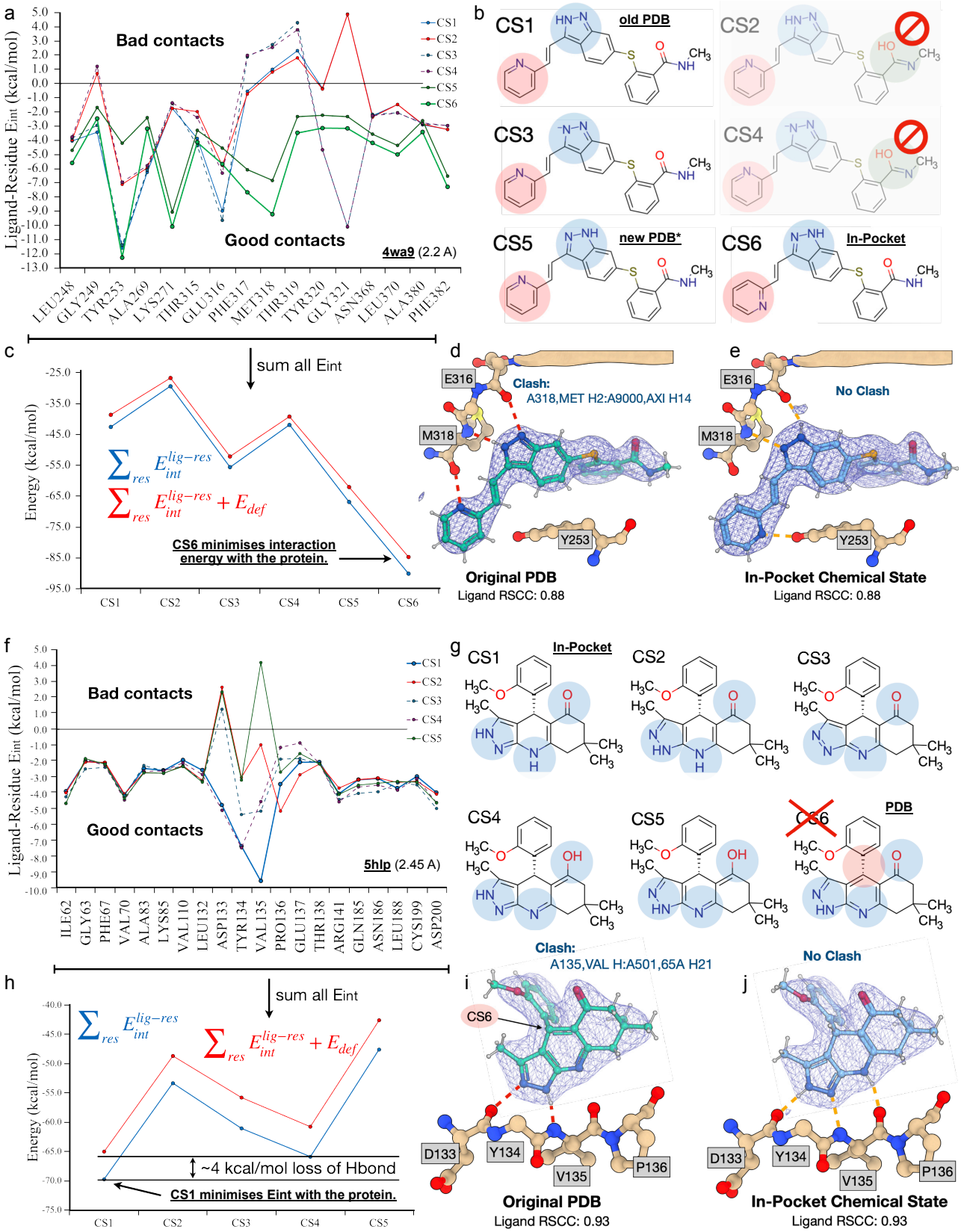


Figure 2: Investigating acid-base properties of ligands in a protein's pocket: correcting ligand chemical states in kinase ATP-pockets. a-e) The binding analysis of axitinib in ABL1. f-j) Pyrazole equilibria in GSK3β. a) The ligand-residue interaction energy

plot for axitinib in ABL1 for six axitinib chemical states identified by IPA. The calculations show that only chemical states CS5 and CS6 avoid suboptimal protein contacts. b) Axitinib chemical states identified by IPA. The chemical states CS2 and CS4 are faded away because these are chemically unrealistic tautomeric states identified by the tautomer search: their energies are approximately 10 kcal/mol higher than the respective amides (CS1 and CS3). The most favourable chemical state is CS6, which should be compared against the present PDB-deposited ligand topology (CS5) and the PDB-deposited chemical state before our reporting (CS1). The circles indicate which ligand functional groups differ between chemical states. c) IPA binding metrics to evaluate the best ligand chemical state in the protein's pocket. The blue curve represents the summed ligand-residue interaction energies. This quantity is corrected in the IPA binding energy by adding the deformation energy (red curve). Irrespective of the metric, IPA selects CS6 as the most reasonable chemical state of axitinib in ABL1's pocket. d,e) Experimental validation: IPA resolves the ligand-protein clash. d) The ligand's tautomeric state before our reporting, and e) the new IPA-identified chemical state. The electron density at the experimental resolution is inconclusive. The depicted model representation also shows why CS6 is more favourable than CS5: by flipping the terminal pyridyl, this group forms an additional hydrogen bond with Y253. g) Ligand-residue interaction energy plot of a pyrazole-containing ligand in GSK3β's pocket for the five IPA-identified ligand chemical states. CS2, CS3, and CS5 lead to poor protein-ligand contacts. h) IPA-identified ligand chemical states (CS1-5) and the PDB-deposited ligand chemical state (CS6). Note that the ligand topology declared in the PDB disagrees with what is reported in the original publication, indicating possible transfer issues. CS6 is disregarded in all calculations due to this inconsistency. i-j) Experimental validation using traditional crystallographic metrics. i) Schematic representation of the pocket and the main hinge residues interacting with the ligand and its density. The structural incompatibility between the PDB-deposited topology and the structural model is evidenced. j) Similar to block i) but for the IPA-selected ligand model. It can be followed that CS1 is more favourable than CS4 due to a hydrogen bond, which yields CS1 an additional 4 kcal/mol stabilisation.

Acid-base equilibria in a protein's pocket.

IPA's most direct application is to determine a ligand's optimal chemical state within a defined protein-ligand binding interface. The algorithm accepts structural models in mmCIF or PDB format as input. After specifying a residue of interest–ligand, amino acid, cofactor, *etc.*, IPA identifies all neighbouring residues within a defined distance threshold–*e.g.*, 4 Å–and it performs an experimentally-guided QM refinement to evaluate ligand-residue binding energetics (**Figure 2a**). Representatively, we chose to first analyse the crystal structure of axitinib bound to unmutated ABL1 (PDB-ID 4wa9; CCD-ID AXI, the ligand identifier).[4] IPA identifies six chemical states, CS1-CS6, which were all considered in the analysis (**Figure 2b**). A fundamental principle for assessing the stability of biological complexes is the absence of ligand-residue pairs with positive interaction energies, as such repulsive contacts likely lead to less favourable total interaction energies and binding (**Figure 2c**). Notably, CS1–corresponding to the previously assigned but, based on this work, corrected PDB ligand topology–displays a proton-proton clash from the 2H-indazole tautomer (**Figure 2d**). Deprotonating the indazole, as in CS3, destabilises interactions with the hinge region (F317-G321) due to unfavourable electronic clashes. Crucially, this effect is only recoverable due to the QM treatment employed by IPA. CS2 and CS4 correspond to high-energy tautomers of CS1 and CS3 and are chemically and biologically irrelevant; only CS5 and CS6 show favourable interaction landscapes among the six identified chemical states, since the 1H-indazole tautomer avoids proton-proton and electron-electron clashes while forming two hydrogen bonds (**Figure 2e**).

At a resolution of 2.2 Å, empirical densities cannot reliably distinguish between carbon and nitrogen atoms (**Figure 2d-e**). As density alone is insufficient for resolving all atomic identities in a structure, IPA detects optimal model interpretations and provides quantitative support for distinguishing relative orientations of functional groups. This underscores the need to examine axitinib's pyridyl group, differentiating CS5 from CS6. Notably, flipping the pyridyl moiety forms an additional, unrecognised hydrogen bond with Y253 (**Figure 2e**), further stabilising the ligand within the

binding pocket. The IPA-identified chemical state, CS6, is further supported by clash analysis (see experimental validation data in **S2**), offering an additional experimental validation of our approach.

In the PDB-ID 5hlp (CCD-ID: 65A),[12] another indazole-containing ligand is bound to the GSK3β kinase (**Figures 2f-j**). The PDB-deposited aromatisation of the central ring is geometrically unrealistic (CS6 in **Figures 2i**): the proposed indazole is unfeasible, so we considered the ligand with a pyrazole group instead. Note that the PDB-deposited ligand topology differs from what is declared in the publication.[12] IPA identifies two ligand chemical states close to the starting PDB topology, with a 1H-pyrazole tautomer: CS2 and CS5. These chemical states clash with the kinase's hinge residues, L132-T138 (**Figures 2i-j**). Deprotonating the pyrazole leads to better interaction networks, as in CS3. However, the ligand's pocket representation is further improved by considering the 2H-pyrazole tautomer, as in CS1 and CS4 (**Figure 2j**). Overall, CS1 is the best chemical representation of this ligand in GSK3β, followed by CS4. Note that though the difference in binding energy is just 4.0 kcal/mol, it is likely that only CS1 is biologically relevant due to an additional hydrogen bond compared to CS4. This assignment agrees with the publication,[12] though a methyl group is still missing from the structure. The crystallographic refinement[13,14] of 5hlp with the IPA-selected chemical state evidences the resolved clash with V135.

Further analyses of ligand-chemical states in protein pockets are provided in the supplements **S2-S6**.

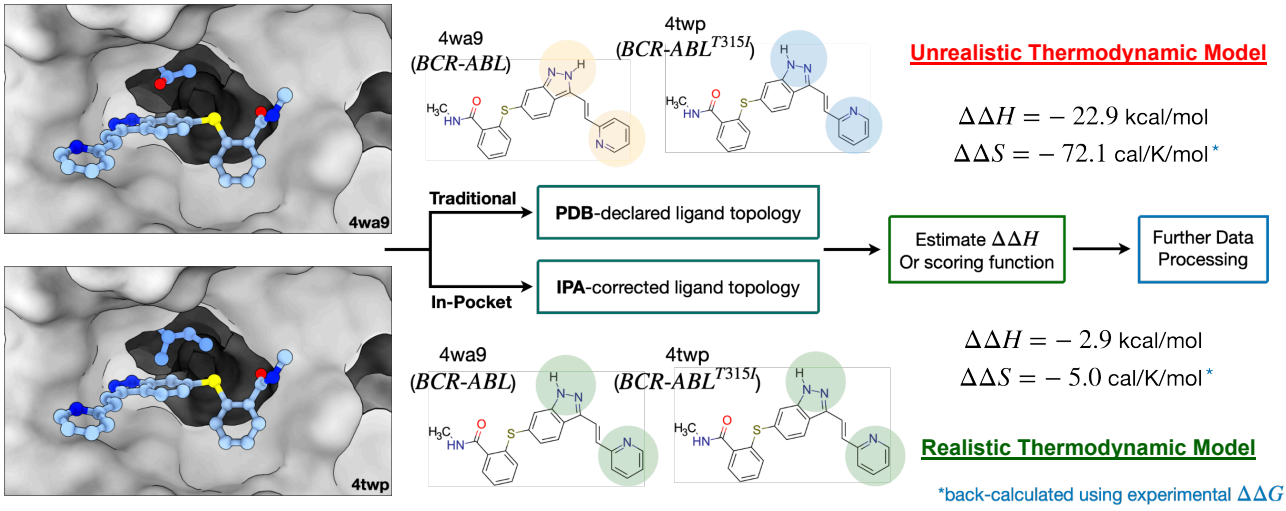


Figure 3: IPA for detailed chemical examination of biological macrocomplexes. Motivational example with two ABL1 variants in complex with axitinib: top left, unmutated protein–PDB-ID 4wa9; bottom left, the T315I mutant–PDB-ID 4twp. The relative QM binding energy should correlate with the relative enthalpy of binding from a hypothetical isothermal titration calorimetry (ITC) experiment. The relative Gibbs free energy is estimated from the IC_{50} ratios, and the relative binding entropies are backcalculated using $\Delta\Delta S = (\Delta\Delta H - \Delta\Delta G)/298.15$. Unrealistic binding thermodynamics data are obtained following the traditional PDB-deposited ligand topology, whereas realistic thermodynamic parameters are obtained using the IPA-suggested ligand chemical state.

The impact of chemical inaccuracies beyond structural biology.

Determining axtinib's chemically plausible state provides its most accurate QM representation, aligns with the crystal structure of ABL1^{T315I} (PDB-ID: 4twp)[4], and yields improved metrics in standard crystallographic validation protocols. We now evaluate the impact of ligand chemical states on thermodynamic predictions by comparing how axitinib engages two ABL1 variants (**Figure 3**). These results highlight how inaccuracies in ligand-state modelling at the protein interface

can propagate through thermodynamic predictions, potentially leading to the misinterpretation of biological data and misleading conclusions in structure-based drug discovery. Two protocols were employed: in one, axitinib was treated as originally deposited in the PDB; in the other, the ligand was modelled using the IPA state. QM energy calculations provide relative binding energies that correlate with binding enthalpies. IPA models produced values closer to those frequently seen in isothermal titration calorimetry experiments, which measure enthalpy and entropy differences. To further validate these findings, we used relative IC_{50} values[4] to estimate the corresponding affinities, from which relative entropies were inferred. These entropy estimates reinforced the observed trends.

Solving ambiguities between structural models.

Multiple structural models may exist for the same protein-ligand complex, either available across different PDB entries or as alternative conformations within a single file. These models may present conflicting representations of ligand binding or simply alternative binding poses in equilibrium with one another. Such is the case for curcumin bound to the DYRK2 kinase: in the PDB-ID 6hdr (CCD-ID CC9), curcumin adopts a binding mode where its central diketone group engages with the kinase's hinge region (**Figure 4a**), whereas in 5ztn (CCD-ID CUR),[15] the diketone moiety is oriented away from the hinge (**Figure 4b**). This discrepancy raises the question of whether one model provides a more accurate description of the biological complex or if both represent equally valid conformations.

Applying IPA, 21 distinct curcumin chemical states were identified (**Figure S7**), all compatible with the experimental data. Each state was evaluated within the three available structural models: one chain from the 6hdr model and two chains (A and B) from 5ztn. Across all cases, IPA consistently identified CS10 as the most stable chemical state in DYRK2's binding pocket (**Figures 4c-e**). This suggests that curcumin's protonation state remains consistent despite the differing ligand binding poses. In this analysis, the IPA binding energy also incorporates the relative energy of each chemical state, accounting for factors such as tautomeric equilibria and potential deprotonation events. Interestingly, CS10 disagrees with the ligand topology proposed in 6hdr–CCD-ID CC9–but aligns with that reported in 5ztn–CCD-ID CUR.

Following this initial assessment, we refined the ligand-residue interaction network to accurately represent hydrogen-bonding patterns, particularly for the phenol groups and the central keto-enol moiety. Equivalent ligand conformations were explored while maintaining the CS10 chemical state across all structural models (**Figures 4f-h, Figure S8**). First, the phenol group adjacent to E193 was reoriented (**Figures 4f-h**, mod1), which reduced the binding energy across all three models (**Figure S8**). Then, a similar rotation was tested for the phenol group near K153 (**Figures 4f-h**, mod2). This stabilised the 5ztn models but was unsuccessful for 6hdr. Finally, recognising that a proton is stabilised between the two oxygens of curcumin's keto-enol group, proton transfers between these oxygens were explored (**Figures 4f-h**, mod3 and mod4; **Figure S8**). This modification lowered the binding energy in 6hdr and 5ztn(A) but not in 5ztn(B).

While these refinements provide a chemically accurate description of curcumin within each structural model, direct inter-model comparisons remain unfeasible because i) the 6hdr structure features a trans-trans curcumin isomer, whereas 5ztn contains a cis-trans isomer, and ii) residual experimental noise varies across models even after IPA refinement. This can be resolved by introducing a renormalisation term, which balances the relative energy of each curcumin-refined model and warrants a direct comparison of all models (Ebind* in **Figure 4i**). This correction strongly disfavours the 5ztn models, ultimately supporting 6hdr as the most chemically realistic representation of curcumin binding to DYRK2. To validate this hypothesis, we incorporated the IPA-refined form of CS10 into 5ztn, adopting a binding mode identical to that

observed in 6hdr (**Figure 4j-k**). This refinement eliminated all ligand-associated clashes and enhanced the model's real-space correlation coefficient (RSCC) score,[16,17] confirming the validity of the IPA model through standard crystallographic metrics.

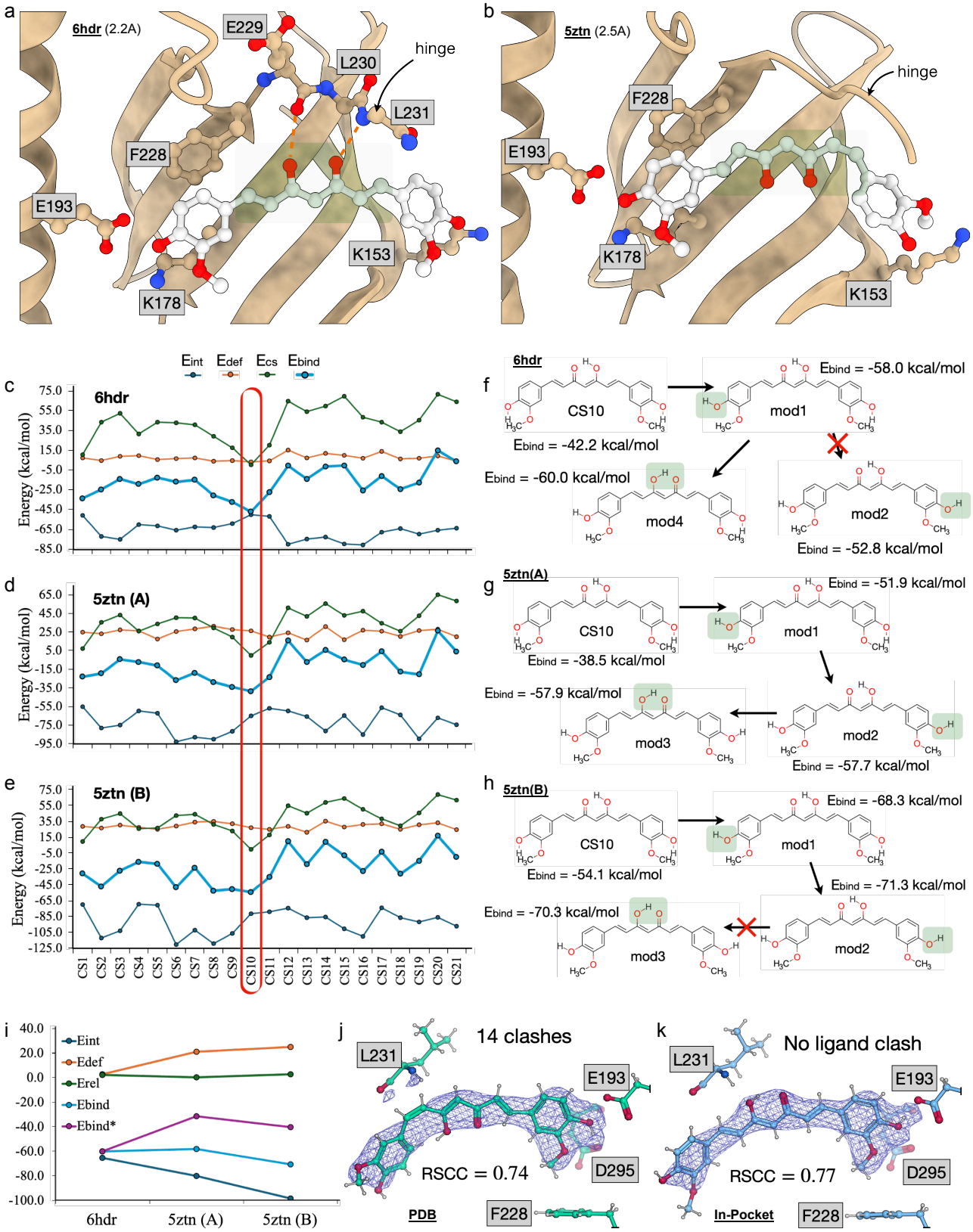


Figure 4: Understanding ambiguities in different protein-ligand structural models and using IPA for an informed selection of the best-representing pocket models: curcumin binding to DYRK2. a) The binding mode of curcumin to DYRK2 in the PDB-ID 6hdr. The two curcumin oxygens from the central diketone/keto-enol scaffold are within the interaction range of the hinge residues E229 and L231. b) The binding mode of curcumin to DYRK2 in the PDB-ID 5ztn, chain A. The central diketone/keto-enol group is flipped when compared to the 6hdr model, indicating that curcumin-hinge interactions are only lipophilic in nature, and no hydrogen bond to the hinge can be formed. There is additionally a change in ligand isomeric state, from trans-trans in 6hdr to trans-cis in 5ztn. c-e) The IPA binding metrics for curcumin in DYRK2. Besides the summed ligand-residue interaction energies and the deformation energy, the IPA binding energy also considers the relative energy of each chemical state to account for slight changes in the relative tautomer energies and the different protonation states. The list of all 21 IPA-identified ligand chemical states is provided in the supplementary material, **Figure S7**. The IPA binding metrics identify CS10 as the most favourable curcumin chemical state in DYRK2's pocket, irrespective of the structural model. f-h) Guided hydrogen-atom network optimisation, as the initial guess proton coordinates might trap the system in metastable local minima: f) The refinement for the 6hdr model, where two changes in the initial proton coordinates are needed; g) The refinement for 5ztn's chain A, where three changes in proton initial coordinates are required; h) The refinement of 5ztn's chain B model, where two changes in the initial proton coordinates are needed. i) Overall IPA binding metrics to compare the structural models 6hdr and 5ztn. To compare the ligand's binding pose in different pockets, a different reference metric is needed, where the change in the isomeric state must be considered. The final metric, Ebind*, represented in purple, clearly shows that 6hdr offers the best pocket representation. j) The 5ztn model validation, as deposited in the PDB. k) Reevaluation of 5ztn's model validation using IPA's selected chemical state and suggested binding mode. The model's fitness has improved.

Typical issues in crystallographic structure solving include transition metal and crystal contact analysis, for which we include in the supplementary material the analysis of hydrolysed imipenem bound to a zinc-containing β-lactamase (PDB-ID 6rzr,[18] **S9**) and of a venetoclax-BCL-2 complex (PDB-ID 6o0k,[19] **S10**).

Mitigating experimental distortion in structural models.

Structure determination pipelines can introduce significant model distortions, impacting the resulting models' accuracy. Representative examples are shown in **Figure 5**. For example, an MDM2 inhibitor (PDB-ID: 4mdn[20], CCD-ID: Y30) exhibits an abnormally distorted formamide group with an ambiguous topology (**Figure 5a**). To assess whether IPA could improve model quality, we refined the structure using the PDB-deposited ligand topology (**Figure 5b**). Initially, the ligand displayed an unrealistically high deformation energy of 1,710.6 kcal/mol, which dropped sharply to 79.7 kcal/mol within a single refinement iteration, highlighting IPA's ability to correct severe structural inaccuracies (**Figure 5c**). Following this initial correction, we pursued two parallel refinement strategies: one optimising the isolated ligand (**Figure 5c**, blue curve) and another refining the ligand within the protein's environment (**Figure 5c**, green curve). The ligand-only refinement resulted in a more pronounced decrease in deformation energy, showing an over-relaxation of the ligand's structure without the protein structural constraints. This effect aligns with expectations, as the flexible IPA bias potential should drive the deformation energy toward zero over many optimisation cycles. To further understand this refinement process, we fitted the deformation energy decay as a function of the number of IPA iterations using power-law functions. The expected deformation energy at 10 IPA-cycles is approximately 35.0 kcal/mol when including the protein, whereas the ligand-only refinement results in values lower by 15.0 kcal/mol. This underscores the importance of accounting for the protein environment in achieving a chemically realistic representation of the ligand within the binding site. Re-running the refinement pipelines led to a minor improvement in RSCC.

We further examined the PDB entry 3zy2 (CCD-ID: GDP)[21] (**Figure 5d**), which contains two alternate conformations of guanosine diphosphate (GDP) bound to the glycosyltransferase POFUT1. Adjacent to GDP's pyrophosphate moiety, a

Mn^{2+} ion coordinates a water molecule. Both alternate ligand conformations exhibit severe distortions (**Figure 5e**), with calculated distortion energies of 712.1 kcal/mol and 585.4 kcal/mol, accounting for the atypical IPA overstabilisation of GDP in the protein's environment (**Figure 5f,g**).

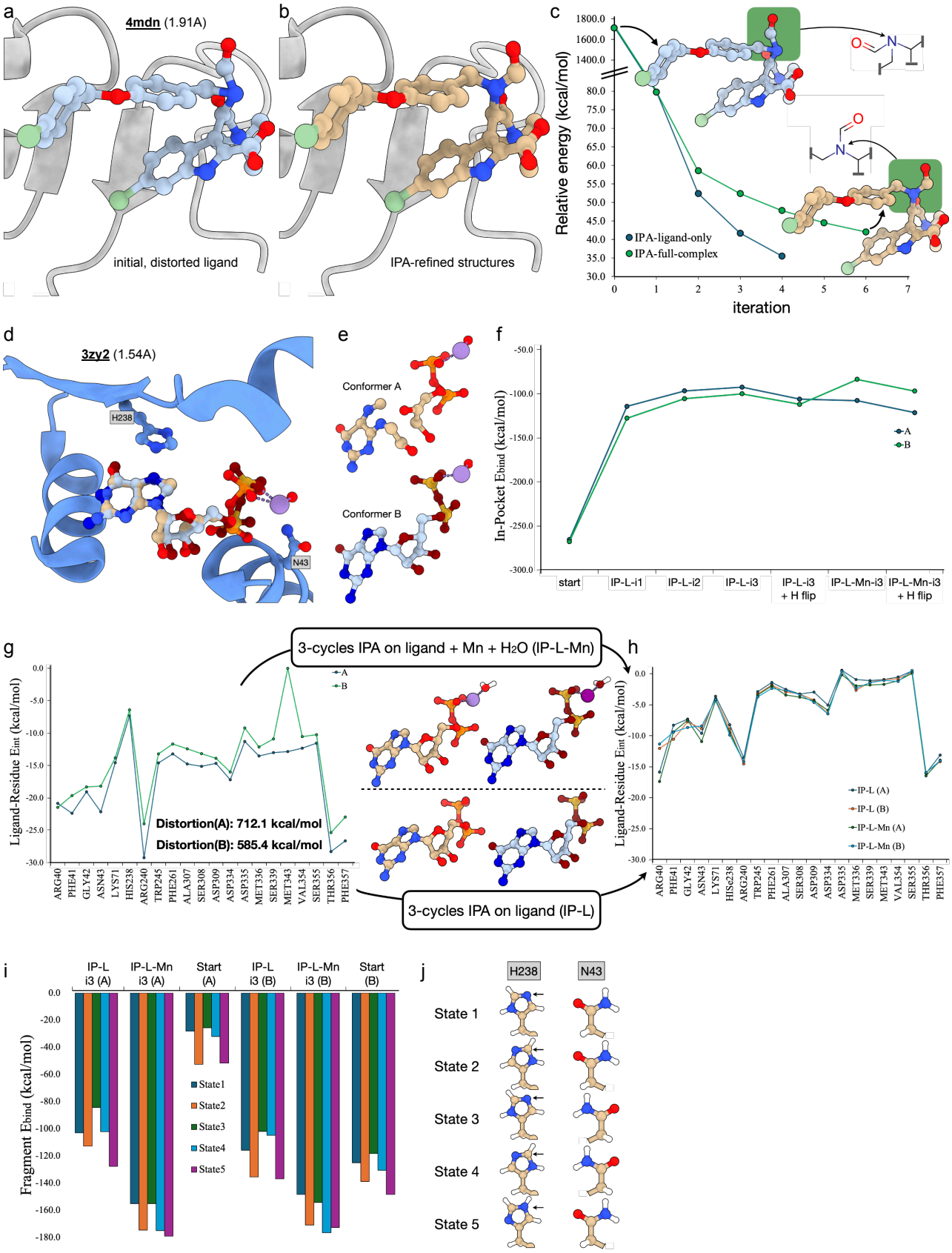


Figure 5: IPA mitigates structural distortions. a) PDB-deposited structure of 4mdn. The formamide group adopts a distorted conformation that deviates from the theoretically expected 120º, resulting from an sp^2 hybridisation. b) The IPA fixed structure of the ligand Y30 shows the formamide group's improved conformation. Two structures are superimposed, one where IPA was applied to the isolated ligand, the other on the whole protein-ligand complex. The structural differences between ligands are minimal. c) Ligand relative energy as a function of IPA refinement iterations. Two refinement strategies were followed after the first ligand-only iteration: we optimised the ligand in isolation (blue) and the ligand within the protein's pocket (green). Though the ligand's geometry is fixed in both cases, we observe a slight drift in the ligand-only optimisation. d) The 3zy2 PDB structure shows two alternate conformations of guanosine diphosphate (GDP) bound to the POFUT1 fucosyltransferase 1. e) The three-dimensional representation of the two alternate conformations of GDP in POFUT1's pocket, showing severe structural distortion. There are bonds missing because the representation software failed to recognise covalent linkages due to the severe structural distortion. f) IPA binding metrics for the two alternate ligand conformations. The ligand's overstabilisation in the pocket observed for the PDB structures results from accumulated distortion. This is significantly improved already with one IPA cycle. The initial PDB structure and the ligand-only IPA-refined structures indicate the degeneracy of the two alternate conformations. The degeneracy is lifted when the Mn^{2+} and the water molecule are included in the IPA ligand refinement cycles. On the abscissa, "start" indicates the initial structure extracted from the PDB; "IP-IL-i1"-"IP-IL-i3" indicates the ligand-only IPA refinement cycles (cycle number indicated last); "IP-IL-i3 + H flip" is identical to "IP-IL-i3" but with flipped H238; the last two points are the result of three IPA refinement cycles where the Mn^{2+} ion was included as part of the ligand, without and with H238 flip. g) IPA's ligand-residue interaction energy plot is severely influenced by the structural distortion of each alternate conformation of GDP in the pocket. The calculated distortion energies are considerable. h) IPA's ligand-residue interaction energy plot after three cycles of IPA ligand refinement using only GDP (path below) and using GDP with a coordinated Mn^{2+} ion and a water molecule bound to the metal (path up). The resulting interaction energy plots are more sensitive and faithful in describing the ligand in the pocket. However, the results are not entirely equivalent. i) The binding energy for fragments extracted from the 3zy2 complex at different residue-orientation states, summarised in j). The lowest binding energies are obtained when the ligand is IPA refined together with the Mn^{2+} ion and the bound water molecule. The results show that the alternate configurations bind preferentially to different protein conformation states.

Given the presence of Mn^{2+}, we tested two three-step IPA curation strategies: one in which only the ligand was refined and another in which Mn^{2+} and its coordinated water molecule were included in the optimisation (**Figure 5g,h**). While both approaches produced comparable results, excluding the metal ion led to a greater structural drift, disrupting the interaction networks of pyrophosphate and Mn^{2+} (*e.g.*, with R40 and K71). This behaviour is expected, given the crucial stabilisation provided by the cation to highly anionic groups like pyrophosphate, reinforcing the requirement for appropriate inclusion of all key chemical-environment elements explicitly, in particular, transition metals and ions.

To quantify the effect of iterative refinements, we plotted the IPA binding energy as a function of refinement cycles (**Figure 5f**). Similar to our previous observations, the most significant distortion correction occurs in the first iteration (IP-L-i1), with progressively minor refinements in subsequent cycles. Notably, the final IPA binding energies of the two ligand-only optimisations converge to the same values, reflecting the structural degeneracy proposed in the original model. We then extended our analysis to three iterations of ligand refinement incorporating Mn^{2+} (IP-L-Mn-i3), expected to yield a more accurate representation of the complex. Interestingly, metal-inclusion splits the total binding energy: conformer A became more stabilised relative to the ligand-only refinement (IP-L-i3), while conformer B was destabilised. This suggests that the PDB-deposited conformational degeneracy may be a model artefact. The effect of H238 protonation was likewise investigated, as default algorithms typically protonate this residue at the δ-position. IPA shows that an ε-protonation is more reasonable, and in all cases, this tautomeric switch lowered the system's energy, supporting the structural model (**Figure 5d**).

Finally, we applied IPA's fragment analysis mode to calculate the binding energies of GDP in complex with the most relevant peptide, metal ion, and water components (**Figure 5i** and **S11**). These calculations, spanning multiple states defined by residue orientations (**Figure 5j**), further reinforced the necessity of including Mn^{2+} in ligand refinement: the lowest binding energies were consistently obtained under the L-Mn refinement protocol, emphasising the metal's critical role in achieving structural completeness. The calculations further support the reorientation of H238, such that its Nε atom becomes a hydrogen bond donor to GDP's N7, allowing H238's Nδ to accept a hydrogen bond from S358.

The refinement pipeline was re-run for the IPA optimised structures, and this model, compared to the original one, improved RSCC by 0.09 units and removed several clashes. **Figure S11** shows IPA evaluations for these models.

Binding sites involving transition metals.

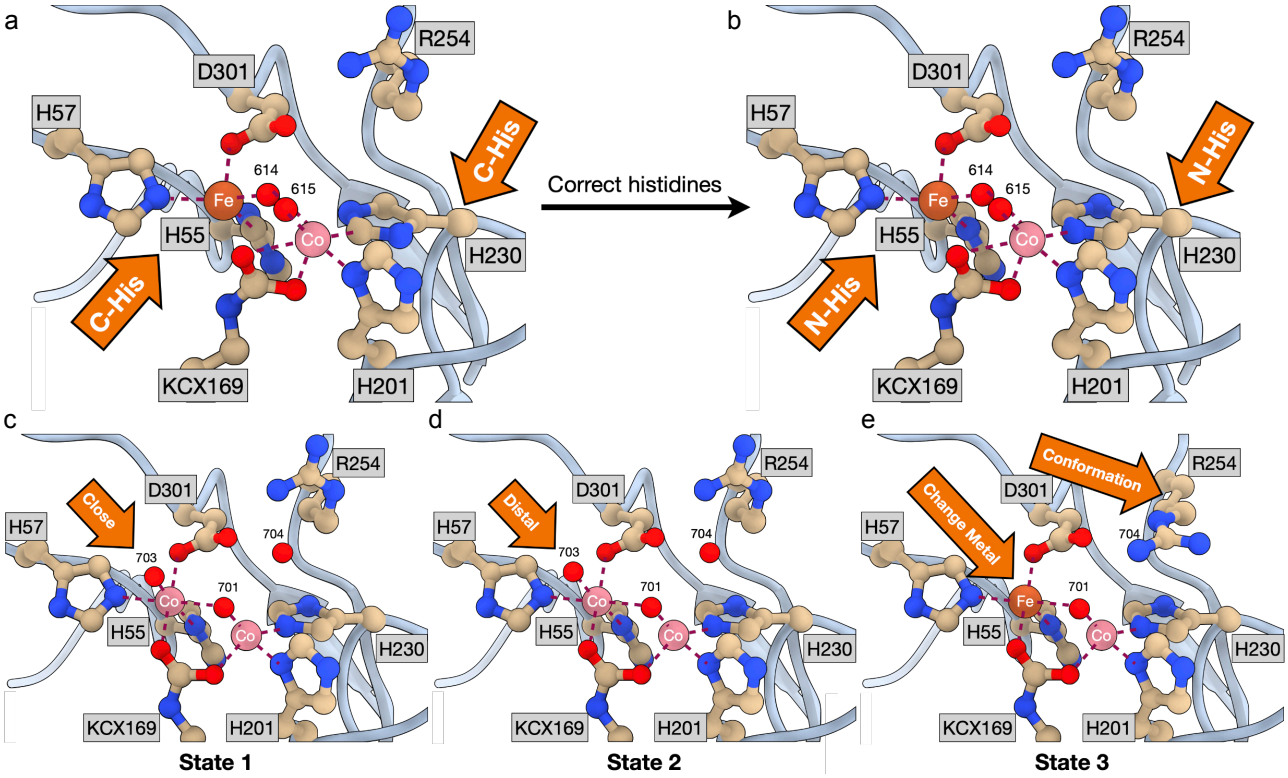


Figure 6: Analysis of the 3wml complex containing two transition metals. a) The PDB-deposited structure of 3wml with two chemically unrealistic C-coordinated histidines. b) After running the IPA protocol, only one conformational state remains, where all metal-coordinated histidine residues do so using either Nε or Nδ. This state avoids metal-proton clashes or the formation of carbenes. The IPA-curated structure was reintroduced in the refinement protocol, from which c-e) three overlapping states were identified. Two states likely contain two cobalt(II) ions, with a bridging water molecule. The two states differ in how water 703 interacts with one of the cobalt centres: c) the water is in the hydroxyl form and coordinated to the metal; d) the water is neutral and uncoordinated. e) The third state likely mixes iron and cobalt. The bridging water remains; however, the iron centre has no equivalent to water 703. The conformation and protonation states of R254 are also changed.

We further applied IPA to a transition metal-containing system. The PDB 3wml[22] features two metal centres–Fe^{2+} and Co^{2+}–coordinated to a post-translationally modified Lys residue (KCX169). In the PDB-deposited model, both metals are coordinated by histidine residues via their Cε atoms (**Figure 6a**). IPA generated sixteen alternative structures by systematically flipping His side chains. Most of these configurations were unstable and led to proton-metal overlaps,

resulting in structural "explosions" (**Figure S12**). As expected, only the configuration featuring four N-coordinated His residues remained reasonable after IPA processing (**Figures 6b** and **S12a**).

The refined structure was reintroduced into the crystallographic refinement protocol following further model building to address areas of positive difference density.[23] This resulted in three distinct, overlapping states (**Figure 6c-e**). Two of these states likely contain two Co^{2+} ions and no Fe^{2+}. A water molecule (HOH701) bridges the two metal centres and, according to IPA, adopts a hydroxyl form. The key difference between these two states lies in how the water molecule HOH703 interacts with the leftmost cobalt ion: in one state, HOH703 coordinates directly to the metal and is also in the hydroxyl state (**Figure 6c**), whereas in the other, the Co–O distance is increased, and IPA identifies HOH703 as an uncoordinated neutral water molecule (**Figure 6d**). The third state (**Figure 6e**) features a mixed-metal configuration with one Fe^{2+} and one Co^{2+} ion, with residue R254 changing its conformational state. Due to the position of H201, R254 must adopt a neutral (deprotonated) state to avoid steric clashes involving its Hε atom. **Figure S13** shows the optimised structures and their calculated relative energies.

DISCUSSION

Structural biology has undergone transformative advances in recent years, driven by breakthrough algorithms built on decades of community effort and the foundational infrastructure of the PDB. Yet, experimental structure determination methods remain fundamentally limited by resolution: many atoms are not directly observed, and atomic identities often cannot be unambiguously assigned. To address these challenges, validation tools have been developed to assess structures against established chemical principles. However, accurately resolving the chemistry within macromolecular binding pockets remained elusive. Misassigned atomic identities lead to flawed interpretations–by both humans and algorithms–with broad downstream consequences for drug design, functional annotation, application of machine learning (ML), and beyond. We show how tailored QM analysis introduces chemical heuristics that augment experimental data, improving model fit and revealing errors otherwise undetectable by conventional methods.

The main focus of this work was the application and validation of IPA using crystallographic data, supported by the underlying experimental data available from the PDB. However, IPA requires only a structural model as input, which can originate from a wide range of sources. For example, in the Supplement S14, IPA is applied to an NMR-derived model, yielding additional refinement beyond the original structure. This allowed fixing, *e.g.*, N-O bonds, which were overly distorted in the initial model. The method can equally be used to evaluate and prioritise AI-predicted models. For instance: i) in the context of ligand validation, as illustrated with the curcumin example discussed above; ii) by identifying the model that exhibits the most comprehensive interaction network, as further explored in the Supplement S15; iii) by refining selected sets of bond distances within a model, as shown in Supplement S16, where all X-H bonds of full proteins were optimised while keeping the experimentally observed coordinates of all heteroatoms. Crucially, owing to its efficient partitioning of the biological system, IPA scales effectively with system size. This is exemplified in the analysis of the cryo-EM model in Supplement S5, where the deposited PDB structure comprises nearly 73,000 atoms.

IPA enables chemically informed validation of structural models. By assessing the plausibility of local configurations, it identifies the most likely chemical states within biological assemblies. Integrating IPA into structure determination pipelines promises to enhance model fidelity–retrospectively, by correcting existing PDB entries, and prospectively, by supporting real-time model assessment. Furthermore, IPA curation is expected to improve the training of structure-aware

protein models by preventing the propagation of chemically implausible configurations that can exacerbate known limitations in predictive accuracy.

METHODS

Theory

QM-based methods have previously been used to enhance the accuracy of protein-ligand structural models. Merz and colleagues introduced a QM/MM refinement protocol integrated into crystallographic workflows.[24] Perola and Charifson,[25] and later Merz and Stewart,[26–28] developed biased methods that constrain energy surfaces with harmonic potentials to align with experimental geometries. Here, we leverage instead an experimentally biased QM framework to determine the most chemically relevant ligand states–protonation states and tautomers–within the environment of a biological macromolecule. This is achieved by restraining the molecular energy surface onto an experimental structure using a metadynamics Gaussian potential:[29]

$$\Delta E = -\kappa N e^{-\alpha \Delta x^2} \tag{1}$$

Δx is the root mean square deviation (RMSD) towards the reference experimental structure, which is the target molecular conformation. κ and α are parameters in the model potential. We consider only positive κ and α, such that the bias potential ΔE is attractive towards the reference structure. N is a factor that we take as the total number of atoms used for calculating the RMSD. This setup permits running constrained minimisations while exploiting the efficiency of unconstrained optimisation algorithms. Ideally, κ should be in the range of 15-30 kcal/mol, whereas α should take values within 0.5 and 1.0 $Å^{-2}$. This allows a rapid and systematic optimisation of structures while retaining conformation: bond distances and angles are the primary variables optimised. This is key to reducing the distortion accumulated due to the limited resolution of experimental techniques.

Using IPA

IPA is available in the form of two scripts. The main script, *In-Pocket-Analysis.py*, is summarised in **Figure S17**. Briefly, the central residue (ligand) is extracted from a PDB or mmCIF file. By default, all residues within 5.0 Å are considered for the calculation. The structures may be protonated using openbabel,[30] RDKit,[31] or a chemical component dictionary (CCD) downloaded from the PDB. The CCD-protonation uses ULYSSES' internal libraries. The search for ligand protonation and tautomeric states is performed with RDKit. The latter is also used to include various amino acid-protonation states. RDKit considers two protonation states for aspartic and glutamic acids, lysines, and tyrosines. For histidines, four protonation states are considered. After each run, an SDF file is generated for the ligand, which may be used to restart calculations. By default, each amino acid is N- and C-capped. We use the closest carbonyl group coordinates and add a hydrogen atom to form an aldehyde (C-terminal) or formamide (N-terminal). The capping effect is also subtracted by default from interaction energies. The program's output includes a text file containing general information (deformation energy, distortion energy, summed interaction energy, relative tautomer energy–if several tautomers are considered, tautomer's total energy, and RMSDs), the ligand-residue interaction energies in a CSV file, all optimised structures, and the outputs of each QM calculation.

The default RDKit run lists all possible chemical states available to the system and found by the algorithm. In some situations, this includes chemical states with negligible biological and chemical relevance, like imidic acid variants of

axitinib (CS2 and CS4 in **Figure 2b**). When running IPA, the user may run calculations on all chemical states or, if prior chemical knowledge is available, specific, preferred ones.

Performance

For a ligand of the size of axitinib with ten chemical states, considering a radius of 4 Å around the central residue (*ca.* 14 residues and their additional protonation states), IPA takes about 3 minutes on a personal laptop.

Availability

IPA is available from the ULYSSES library.[32] A stand-alone tool will be available in https://github.com/s1riius/In-Pocket-Analysis.git upon publication. We are currently also integrating IPA into CCP4.[33] Note that, due to the specificities of some licenses, not all protonation engines might be available from all interfaces.

Limitations

The primary limitation of the IPA code lies in its treatment of only pairwise (residue-residue) interactions, which often fall short of capturing the full complexity of biological environments. Notably, this includes regions with localised accumulation of carboxylate groups or coordination spheres involving transition metals–situations where higher-order interactions are critical. To address such cases, we introduce the IPA-fragment mode, which was essential in several parts of this study. This extension enables the treatment of more chemically complex environments, though at a higher computational cost: the scaling increases approximately cubically with system size. Users are therefore advised to account for this when setting up systems for further calculations.

Finally, while IPA-derived binding energies offer valuable insight into the chemistry of binding pockets, caution is warranted when applying them as scoring functions, for instance, in drug discovery workflows. Although such use is briefly explored in this work, broader validation is needed before routine application in this context.

Data Availability

All the IPA-optimised ligands from the PDBbind[8]/MISATO[9] databases are available in https://gitlab.com/siriius/ipa-ligands.git. Here, QM-refined geometries and the respective molecular properties are available for download.

ACKNOWLEDGEMENTS

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF), projects SUPREME (number: 031L0268) and DATIPilot Sprint – MISATO (03DPS1234); by the Bundesministerium für Wirtschaft und Klimaschutz (BMWK) via ZIM (grant number: KK5710001BA4); by the Deutsche Forschungsgemeinschaft, DFG grant CRC387, project ID TRR 387/1-514894665.

COMPETING INTERESTS

The authors declare no competing interests, financial or otherwise.

REFERENCES

(1) wwPDB consortium; Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. D.; Christie, C.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson,

B. P.; Liang, Y.; Lowe, R.; Peisach, E.; Periskova, I.; Randle, C.; Rose, A.; Sekharan, M.; Shao, C.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Young, J.; Zardecki, C.; Zhuravleva, M.; Kurisu, G.; Nakamura, H.; Kengaku, Y.; Cho, H.; Sato, J.; Kim, J. Y.; Ikegawa, Y.; Nakagawa, A.; Yamashita, R.; Kudou, T.; Bekker, G.-J.; Suzuki, H.; Iwata, T.; Yokochi, M.; Kobayashi, N.; Fujiwara, T.; Velankar, S.; Kleywegt, G. J.; Anyango, S.; Armstrong, D. R.; Berrisford, J. M.; Conroy, M. J.; Dana, J. M.; Deshpande, M.; Gane, P.; Gáborová, R.; Gupta, D.; Gutmanas, A.; Koča, J.; Mak, L.; Mir, S.; Mukhopadhyay, A.; Nadzirin, N.; Nair, S.; Patwardhan, A.; Paysan-Lafosse, T.; Pravda, L.; Salih, O.; Sehnal, D.; Varadi, M.; Vařeková, R.; Markley, J. L.; Hoch, J. C.; Romero, P. R.; Baskaran, K.; Maziuk, D.; Ulrich, E. L.; Wedell, J. R.; Yao, H.; Livny, M.; Ioannidis, Y. E. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47* (D1), D520–D528. https://doi.org/10.1093/nar/gky949.

(2) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(3) Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; McHugh, R.; Vafeados, D.; Li, X.; Sutherland, G. A.; Hitchcock, A.; Hunter, C. N.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Baek, M.; DiMaio, F.; Baker, D. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *Science* **2024**, *384* (6693), eadl2528. https://doi.org/10.1126/science.adl2528.

(4) Pemovska, T.; Johnson, E.; Kontro, M.; Repasky, G. A.; Chen, J.; Wells, P.; Cronin, C. N.; McTigue, M.; Kallioniemi, O.; Porkka, K.; Murray, B. W.; Wennerberg, K. Axitinib Effectively Inhibits BCR-ABL1(T315I) with a Distinct Binding Conformation. *Nature* **2015**, *519* (7541), 102–105. https://doi.org/10.1038/nature14119.

(5) Kühlbrandt, W. The Resolution Revolution. *Science* **2014**, *343* (6178), 1443–1444. https://doi.org/10.1126/science.1251652.

(6) Steiner, R. A.; Tucker, J. A. Keep It Together: Restraints in Crystallographic Refinement of Macromolecule–Ligand Complexes. *Acta Crystallogr. Sect. Struct. Biol.* **2017**, *73* (2), 93–102. https://doi.org/10.1107/s2059798316017964.

(7) Cruickshank, D. W. J. Remarks about Protein Structure Precision. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55* (3), 583–601. https://doi.org/10.1107/s0907444998012645.

(8) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

(9) Siebenmorgen, T.; Menezes, F.; Benassou, S.; Merdivan, E.; Didi, K.; Mourão, A. S. D.; Kitel, R.; Liò, P.; Kesselheim, S.; Piraud, M.; Theis, F. J.; Sattler, M.; Popowicz, G. M. MISATO: Machine Learning Dataset of Protein–Ligand Complexes for Structure-Based Drug Discovery. *Nat. Comput. Sci.* **2024**, *4* (5), 367–378. https://doi.org/10.1038/s43588-024-00627-2.

(10) *NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 22, May 2022, Editor: Russell D. Johnson III http://cccbdb.nist.gov/.*

(11) Dorofeeva, O. V.; Vishnevskiy, Y. V.; Vogt, N.; Vogt, J.; Khristenko, L. V.; Krasnoshchekov, S. V.; Shishkov, I. F.; Hargittai, I.; Vilkov, L. V. Molecular Structure and Conformation of Nitrobenzene Reinvestigated by Combined Analysis of Gas-Phase Electron Diffraction, Rotational Constants, and Theoretical Calculations. *Struct. Chem.* **2007**, *18* (6), 739–753. https://doi.org/10.1007/s11224-007-9186-6.

(12) Wagner, F. F.; Bishop, J. A.; Gale, J. P.; Shi, X.; Walk, M.; Ketterman, J.; Patnaik, D.; Barker, D.; Walpita, D.; Campbell, A. J.; Nguyen, S.; Lewis, M.; Ross, L.; Weïwer, M.; An, W. F.; Germain, A. R.; Nag, P. P.; Metkar, S.; Kaya, T.; Dandapani, S.; Olson, D. E.; Barbe, A.-L.; Lazzaro, F.; Sacher, J. R.; Cheah, J. H.; Fei, D.; Perez, J.; Munoz, B.; Palmer, M.; Stegmaier, K.; Schreiber, S. L.; Scolnick, E.; Zhang, Y.-L.; Haggarty, S. J.; Holson, E. B.; Pan, J. Q. Inhibitors of Glycogen Synthase Kinase 3 with Exquisite Kinome-Wide Selectivity and Their Functional Effects. *ACS Chem. Biol.* **2016**, *11* (7), 1952–1963. https://doi.org/10.1021/acschembio.6b00306.

(13) Yamashita, K.; Wojdyr, M.; Long, F.; Nicholls, R. A.; Murshudov, G. N. *GEMMI* and *Servalcat* Restrain *REFMAC*5. *Acta Crystallogr. Sect. Struct. Biol.* **2023**, *79* (5), 368–373. https://doi.org/10.1107/s2059798323002413.

(14) Yamashita, K.; Palmer, C. M.; Burnley, T.; Murshudov, G. N. Cryo-EM Single-Particle Structure Refinement and Map Calculation Using*Servalcat*. *Acta Crystallogr. Sect. Struct. Biol.* **2021**, *77* (10), 1282–1291. https://doi.org/10.1107/s2059798321009475.

(15) Banerjee, S.; Ji, C.; Mayfield, J. E.; Goel, A.; Xiao, J.; Dixon, J. E.; Guo, X. Ancient Drug Curcumin Impedes 26S Proteasome Activity by Direct Inhibition of Dual-Specificity Tyrosine-Regulated Kinase 2. *Proc. Natl. Acad. Sci.* **2018**, *115* (32), 8155–8160. https://doi.org/10.1073/pnas.1806797115.

(16) *https://github.com/project-gemmi/gemmi*.

(17) Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. The*PDB_REDO*Server for Macromolecular Structure Model Optimization. *IUCrJ* **2014**, *1* (4), 213–220. https://doi.org/10.1107/s2052252514009324.
(18) Softley, C. A.; Zak, K. M.; Bostock, M. J.; Fino, R.; Zhou, R. X.; Kolonko, M.; Mejdi-Nitiu, R.; Meyer, H.; Sattler, M.; Popowicz, G. M. Structure and Molecular Recognition Mechanism of IMP-13 Metallo-β-Lactamase. *Antimicrob. Agents Chemother.* **2020**, *64* (6), e00123-20. https://doi.org/10.1128/AAC.00123-20.
(19) Birkinshaw, R. W.; Gong, J.; Luo, C. S.; Lio, D.; White, C. A.; Anderson, M. A.; Blombery, P.; Lessene, G.; Majewski, I. J.; Thijssen, R.; Roberts, A. W.; Huang, D. C. S.; Colman, P. M.; Czabotar, P. E. Structures of BCL-2 in Complex with Venetoclax Reveal the Molecular Basis of Resistance Mutations. *Nat. Commun.* **2019**, *10* (1), 2385. https://doi.org/10.1038/s41467-019-10363-1.
(20) Bista, M.; Wolf, S.; Khoury, K.; Kowalska, K.; Huang, Y.; Wrona, E.; Arciniega, M.; Popowicz, G. M.; Holak, T. A.; Dömling, A. Transient Protein States in Designing Inhibitors of the MDM2-P53 Interaction. *Structure* **2013**, *21* (12), 2143–2151. https://doi.org/10.1016/j.str.2013.09.006.
(21) Lira-Navarrete, E.; Valero-González, J.; Villanueva, R.; Martínez-Júlvez, M.; Tejero, T.; Merino, P.; Panjikar, S.; Hurtado-Guerrero, R. Structural Insights into the Mechanism of Protein O-Fucosylation. *PLoS ONE* **2011**, *6* (9), e25365. https://doi.org/10.1371/journal.pone.0025365.
(22) Naqvi, T.; Warden, A. C.; French, N.; Sugrue, E.; Carr, P. D.; Jackson, C. J.; Scott, C. A 5000-Fold Increase in the Specificity of a Bacterial Phosphotriesterase for Malathion through Combinatorial Active Site Mutagenesis. *PLoS ONE* **2014**, *9* (4), e94177. https://doi.org/10.1371/journal.pone.0094177.
(23) Babai, K. H.; Long, F.; Malý, M.; Yamashita, K.; Murshudov, G. N. Improving Macromolecular Structure Refinement with Metal-Coordination Restraints. *Acta Crystallogr. Sect. Struct. Biol.* **2024**, *80* (12), 821–833. https://doi.org/10.1107/S2059798324011458.
(24) Fu, Z.; Li, X.; Merz, K. M. Accurate Assessment of the Strain Energy in a Protein-bound Drug Using QM/MM X-ray Refinement and Converged Quantum Chemistry. *J. Comput. Chem.* **2011**, *32* (12), 2587–2597. https://doi.org/10.1002/jcc.21838.
(25) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47* (10), 2499–2510. https://doi.org/10.1021/jm030563w.
(26) Yu, N.; Yennawar, H. P.; Merz, K. M. Refinement of Protein Crystal Structures Using Energy Restraints Derived from Linear-Scaling Quantum Mechanics. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (3), 322–332. https://doi.org/10.1107/S0907444904033669.
(27) Yu, N.; Li, X.; Cui, G.; Hayik, S. A.; Merz, K. M. Critical Assessment of Quantum Mechanics Based Energy Restraints in Protein Crystal Structure Refinement. *Protein Sci.* **2006**, *15* (12), 2773–2784. https://doi.org/10.1110/ps.062343206.
(28) Stewart, J. J. P. Application of the PM6 Method to Modeling Proteins. *J. Mol. Model.* **2009**, *15* (7), 765–805. https://doi.org/10.1007/s00894-008-0420-y.
(29) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.
(30) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.
(31) *RDKit: Open-source cheminformatics. https://www.rdkit.org*.
(32) Menezes, F.; Popowicz, G. M. *ULYSSES* : An Efficient and Easy to Use Semiempirical Library for C++. *J. Chem. Inf. Model.* **2022**, *62* (16), 3685–3694. https://doi.org/10.1021/acs.jcim.2c00757.
(33) Agirre, J.; Atanasova, M.; Bagdonas, H.; Ballard, C. B.; Baslé, A.; Beilsten-Edmands, J.; Borges, R. J.; Brown, D. G.; Burgos-Mármol, J. J.; Berrisford, J. M.; Bond, P. S.; Caballero, I.; Catapano, L.; Chojnowski, G.; Cook, A. G.; Cowtan, K. D.; Croll, T. I.; Debreczeni, J. É.; Devenish, N. E.; Dodson, E. J.; Drevon, T. R.; Emsley, P.; Evans, G.; Evans, P. R.; Fando, M.; Foadi, J.; Fuentes-Montero, L.; Garman, E. F.; Gerstel, M.; Gildea, R. J.; Hatti, K.; Hekkelman, M. L.; Heuser, P.; Hoh, S. W.; Hough, M. A.; Jenkins, H. T.; Jiménez, E.; Joosten, R. P.; Keegan, R. M.; Keep, N.; Krissinel, E. B.; Kolenko, P.; Kovalevskiy, O.; Lamzin, V. S.; Lawson, D. M.; Lebedev, A. A.; Leslie, A. G. W.; Lohkamp, B.; Long, F.; Malý, M.; McCoy, A. J.; McNicholas, S. J.; Medina, A.; Millán, C.; Murray, J. W.; Murshudov, G. N.; Nicholls, R. A.; Noble, M. E. M.; Oeffner, R.; Pannu, N. S.; Parkhurst, J. M.; Pearce, N.; Pereira, J.; Perrakis, A.; Powell, H. R.; Read, R. J.; Rigden, D. J.; Rochira, W.; Sammito, M.; Sánchez Rodríguez, F.; Sheldrick, G. M.; Shelley, K. L.; Simkovic, F.; Simpkin, A. J.; Skubak, P.; Sobolev, E.; Steiner, R. A.; Stevenson, K.; Tews, I.; Thomas, J. M. H.; Thorn, A.; Valls, J. T.; Uski, V.; Usón, I.; Vagin, A.; Velankar, S.; Vollmar, M.; Walden, H.; Waterman, D.; Wilson, K. S.; Winn, M. D.; Winter, G.; Wojdyr, M.; Yamashita, K. The *CCP* 4 Suite: Integrative Software for Macromolecular Crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **2023**, *79* (6), 449–461. https://doi.org/10.1107/S2059798323003595.
(34) Demuro, S.; Sauvey, C.; Tripathi, S. K.; Di Martino, R. M. C.; Shi, D.; Ortega, J. A.; Russo, D.; Balboni, B.; Giabbai, B.; Storici, P.; Girotto, S.; Abagyan, R.; Cavalli, A. ARN25068, a Versatile Starting Point towards Triple GSK-3β/FYN/DYRK1A Inhibitors to Tackle Tau-Related Neurological Disorders. *Eur. J. Med. Chem.* **2022**, *229*, 114054. https://doi.org/10.1016/j.ejmech.2021.114054.

(35) Bosanac, T.; Hickey, E. R.; Ginn, J.; Kashem, M.; Kerr, S.; Kugler, S.; Li, X.; Olague, A.; Schlyer, S.; Young, E. R. R. Substituted 2H-Isoquinolin-1-Ones as Potent Rho-Kinase Inhibitors: Part 3, Aryl Substituted Pyrrolidines. *Bioorg. Med. Chem. Lett.* **2010**, *20* (12), 3746–3749. https://doi.org/10.1016/j.bmcl.2010.04.069.
(36) Sang Jae; Choi, J.-S.; Bong, S. M.; Hwang, H.-J.; Lee, J.; Song, H.-J.; Lee, J.; Kim, J.-H.; Koh, J. S.; Lee, B. I. Crystal Structures of Spleen Tyrosine Kinase in Complex with Two Novel 4-Aminopyrido[4,3-d] Pyrimidine Derivative Inhibitors. *Mol. Cells* **2018**, *41* (6), 545–552. https://doi.org/10.14348/molcells.2018.2219.
(37) Hochheiser, I. V.; Pilsl, M.; Hagelueken, G.; Moecking, J.; Marleaux, M.; Brinkschulte, R.; Latz, E.; Engel, C.; Geyer, M. Structure of the NLRP3 Decamer Bound to the Cytokine Release Inhibitor CRID3. *Nature* **2022**, *604* (7904), 184–189. https://doi.org/10.1038/s41586-022-04467-w.
(38) Caine, B. A.; Bronzato, M.; Popelier, P. L. A. Experiment Stands Corrected: Accurate Prediction of the Aqueous p K_a Values of Sulfonamide Drugs Using Equilibrium Bond Lengths. *Chem. Sci.* **2019**, *10* (25), 6368–6381. https://doi.org/10.1039/C9SC01818B.
(39) Liu, F.; Wang, B.; Wang, Q.; Qi, Z.; Chen, C.; Kong, L.-L.; Chen, J.-Y.; Liu, X.; Wang, A.; Hu, C.; Wang, W.; Wang, H.; Wu, F.; Ruan, Y.; Qi, S.; Liu, J.; Zou, F.; Hu, Z.; Wang, W.; Wang, L.; Zhang, S.; Yun, C.-H.; Zhai, Z.; Liu, J.; Liu, Q. Discovery and Characterization of a Novel Potent Type II Native and Mutant BCR-ABL Inhibitor (CHMFL-074) for Chronic Myeloid Leukemia (CML). *Oncotarget* **2016**, *7* (29), 45562–45574. https://doi.org/10.18632/oncotarget.10037.
(40) Zeng, L.; Li, J.; Muller, M.; Yan, S.; Mujtaba, S.; Pan, C.; Wang, Z.; Zhou, M.-M. Selective Small Molecules Blocking HIV-1 Tat and Coactivator PCAF Association. *J. Am. Chem. Soc.* **2005**, *127* (8), 2376–2377. https://doi.org/10.1021/ja044885g.
(41) Alzari, P. M.; Ne Souchon, H.; Dominguez, R. The Crystal Structure of Endoglucanase CelA, a Family 8 Glycosyl Hydrolase from Clostridium Thermocellum. *Structure* **1996**, *4* (3), 265–275. https://doi.org/10.1016/S0969-2126(96)00031-7.
(42) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and Better Reference Data for Improved All-atom Structure Validation. *Protein Sci.* **2018**, *27* (1), 293–315. https://doi.org/10.1002/pro.3330.
(43) Hekkelman, M. L.; Te Beek, T. A. H.; Pettifer, S. R.; Thorne, D.; Attwood, T. K.; Vriend, G. WIWS: A Protein Structure Bioinformatics Web Service Collection. *Nucleic Acids Res.* **2010**, *38* (Web Server), W719–W723. https://doi.org/10.1093/nar/gkq453.
(44) Yagi, D.; Yamada, T.; Kurihara, K.; Ohnishi, Y.; Yamashita, M.; Tamada, T.; Tanaka, I.; Kuroki, R.; Niimura, N. A Neutron Crystallographic Analysis of Phosphate-Free Ribonuclease A at 1.7 Å Resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2009**, *65* (9), 892–899. https://doi.org/10.1107/S0907444909018885.
(45) Dawidowski, M.; Emmanouilidis, L.; Kalel, V. C.; Tripsianes, K.; Schorpp, K.; Hadian, K.; Kaiser, M.; Mäser, P.; Kolonko, M.; Tanghe, S.; Rodriguez, A.; Schliebs, W.; Erdmann, R.; Sattler, M.; Popowicz, G. M. Inhibitors of PEX14 Disrupt Protein Import into Glycosomes and Kill *Trypanosoma* Parasites. *Science* **2017**, *355* (6332), 1416–1420. https://doi.org/10.1126/science.aal1807.
(46) Gardberg, A. S.; Del Castillo, A. R.; Weiss, K. L.; Meilleur, F.; Blakeley, M. P.; Myles, D. A. A. Unambiguous Determination of H-Atom Positions: Comparing Results from Neutron and High-Resolution X-Ray Crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (5), 558–567. https://doi.org/10.1107/S0907444910005494.
(47) Chen, J. C.-H.; Hanson, B. L.; Fisher, S. Z.; Langan, P.; Kovalevsky, A. Y. Direct Observation of Hydrogen Atom Dynamics and Interactions by Ultrahigh Resolution Neutron Protein Crystallography. *Proc. Natl. Acad. Sci.* **2012**, *109* (38), 15301–15306. https://doi.org/10.1073/pnas.1208341109.
(48) Park, J.; Selvaraj, B.; McShan, A. C.; Boyken, S. E.; Wei, K. Y.; Oberdorfer, G.; DeGrado, W.; Sgourakis, N. G.; Cuneo, M. J.; Myles, D. A.; Baker, D. De Novo Design of a Homo-Trimeric Amantadine-Binding Protein. *eLife* **2019**, *8*, e47839. https://doi.org/10.7554/eLife.47839.

It's Not Always Crystal Clear: Recovering Chemical Dynamics Using In-Pocket Analysis of Protein-Ligand Binding

SUPPORTING INFORMATION

S1 – The relationship between distortion and deformation

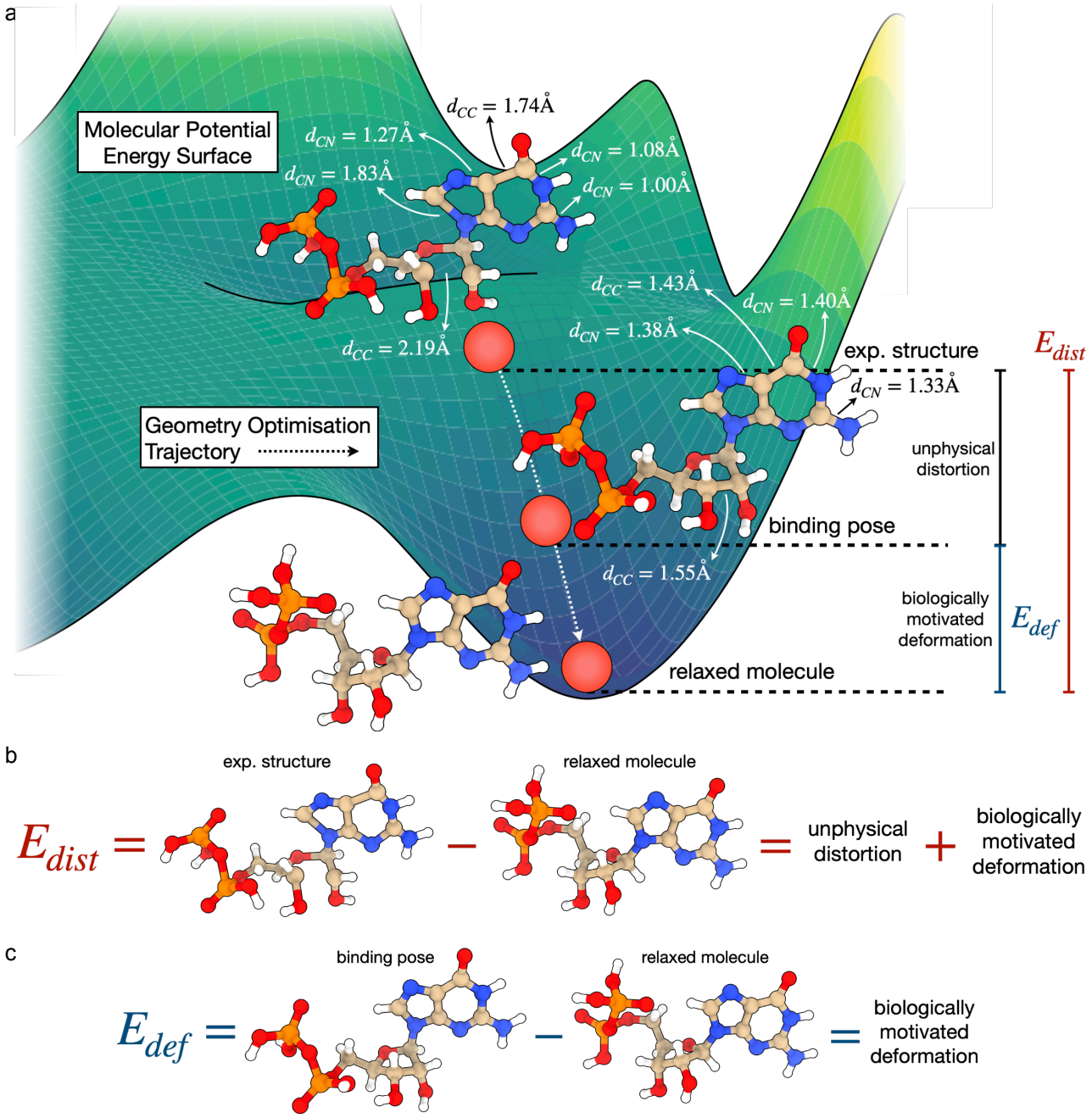


Figure S1: Theoretical background depicting the rationale behind IPA and how distortion and deformation correlate with one another. a) Schematic representation of the various ligand conformational states on the ligand-only potential energy surface (PES). The experimentally observed ligand structure ("exp. structure") is depicted as the red sphere at the top, corresponding to the molecular geometry shown above it. This structure includes biologically meaningful distortions induced by the biomolecular target, as well as unphysical distortions resulting from limited

experimental resolution and modelling inaccuracies (*e.g.*, incorrectly applying restraints while solving a structure). Note that unphysical distortion may arise for lower resolution structures because then the structural model becomes heavily reliant on correctly using and the validity of distance and angular restraints. Examples of such inaccuracies include abnormal bond lengths and angles present in the PDB-deposited model, but can also include atomic clashes, *e.g.*, by refining structures in the absence of hydrogen atoms. In this illustration, we imagine performing a structural relaxation of the ligand in the absence of the protein. The geometry optimisation trajectory is assumed to initially correct the inaccurate bond lengths and angles while maintaining the ligand's binding pose conformation. This stage primarily removes unphysical distortions from the structure. The trajectory then proceeds to fully refine the ligand's geometry, relaxing torsional angles and, therefore, also the binding pose. Ultimately, this yields a representation of the free ligand in solution, absent from the influence of the biomolecular environment. b) Naïve calculation of deformation energies. This approach uses the processed experimental structure from the PDB model and compares its energy to that of the fully relaxed ligand, in this case, QM optimised. As illustrated in a), this energy difference includes not only the biologically relevant deformation but also unphysical distortions. We refer to this energy as the distortion energy (E_{dist}). c) Biologically meaningful calculation of deformation energies (E_{def}), *i.e.*, including exclusively the protein-induced change in ligand conformation. This method uses chemically reasonable representations of the ligand–its binding poses without unphysical distortions in bond lengths and angles–and compares them to the fully relaxed, quantum mechanically optimised structure. In this work, we introduce IPA optimisation, a method capable of refining ligand structures while preserving their initial binding conformations. This enables structural relaxation in the absence of the protein binding site. According to our definitions, E_{dist} encompasses E_{def}–the biologically meaningful deformation energy–along with additional, unwanted structural distortions resulting from limited resolution of the experimental data. While one could define E_{dist} to include only these unphysical distortions, doing so would not reflect how deformation energies are typically calculated in practice, nor would it allow for a meaningful comparison between E_{dist} and E_{def}. The deformation energy is an important quantity to estimate to study biomolecular binding, as this quantity has a negative impact on binding thermodynamics. If ligand deformation energies are unrealistically too large, binding cannot take place. It is therefore important to correctly estimate E_{def} in order to obtain a more realistic representation of a protein-ligand complex.

S2 – Other chain data for 4wa9 and 5hlp

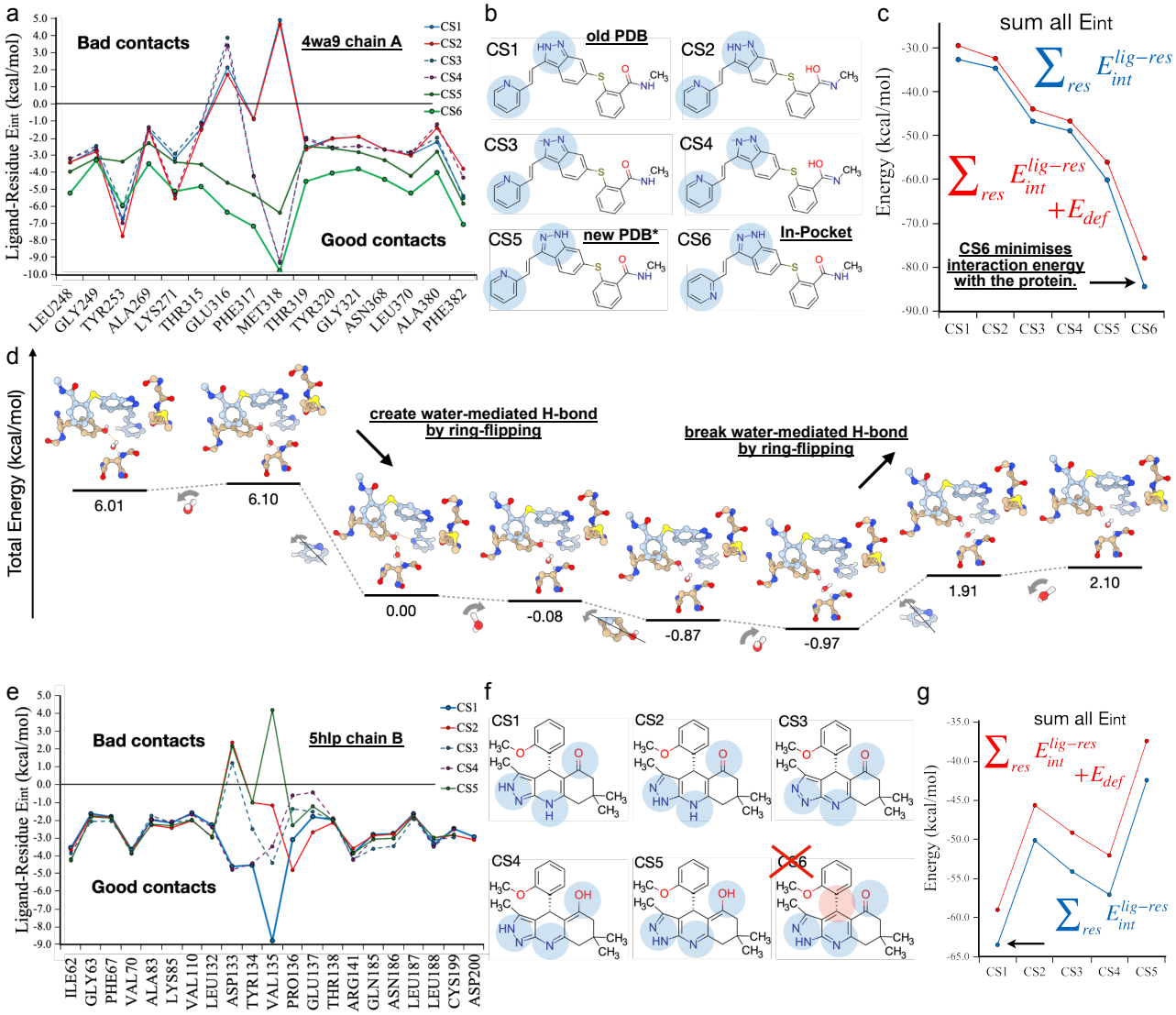


Figure S2: IPA data for other chains of PDB-IDs 4wa9 (chain A, blocks a-d) and 5hlp (chain B, blocks e-g). a) The ligand-residue interaction energy plot shows that CS5 and CS6 are the only two chemical states that do not clash with the kinase's hinge. Of these two CSs, CS6 offers predominantly better interactions with Y253 and M318. Note that although CS5 and CS6 differ only by a 180º flip of the terminal pyridyl, this leads to slight differences in the IPA refined ligand structures. These fine differences justify the observed shift between the CS5 and CS6 curves. Interestingly, the energy gain by flipping the terminal pyridyl from CS5 to CS6 regarding the interaction with Y253 is not as pronounced as for chain B. b) The chemical states identified by IPA, with structural changes highlighted. c) The summed interaction energies and the total IPA binding energy (summed interaction energies and the deformation energy) show that CS6 is still the most stabilised axitinib chemical state in ABL1's pocket. d) A fragment interaction energy plot showing that the pyridyl flipping stabilisation with Y253 is, in fact, a water-mediated hydrogen bond for chain A. This shows that the enthalpy stabilisation with Y253 is partly transferred to a water-mediated interaction. e) The ligand-residue interaction energy plot for 5hlp, chain B. Despite minor numerical differences, the results fully agree with those for chain A. f) The ligand chemical states identified by IPA, with differences highlighted in blue or red circles. Note that the PDB-deposited topology is inconsistent with the available structural model. g) The summed interaction energies and the IPA binding

energy for the different chemical states show that CS1 is the most stable chemical state of this ligand in GSK3β's pocket. This is followed by CS4. The difference in binding energy also reflects the loss of an H-bond.

Experimental data: For accessing all the data underlying the results and the experimental validation using traditional crystallographic metrics, check the "supplements/figure2+S1/" folder, followed by the PDB-IDs.

S3 – An additional pyrazole-indazole chemical equilibrium

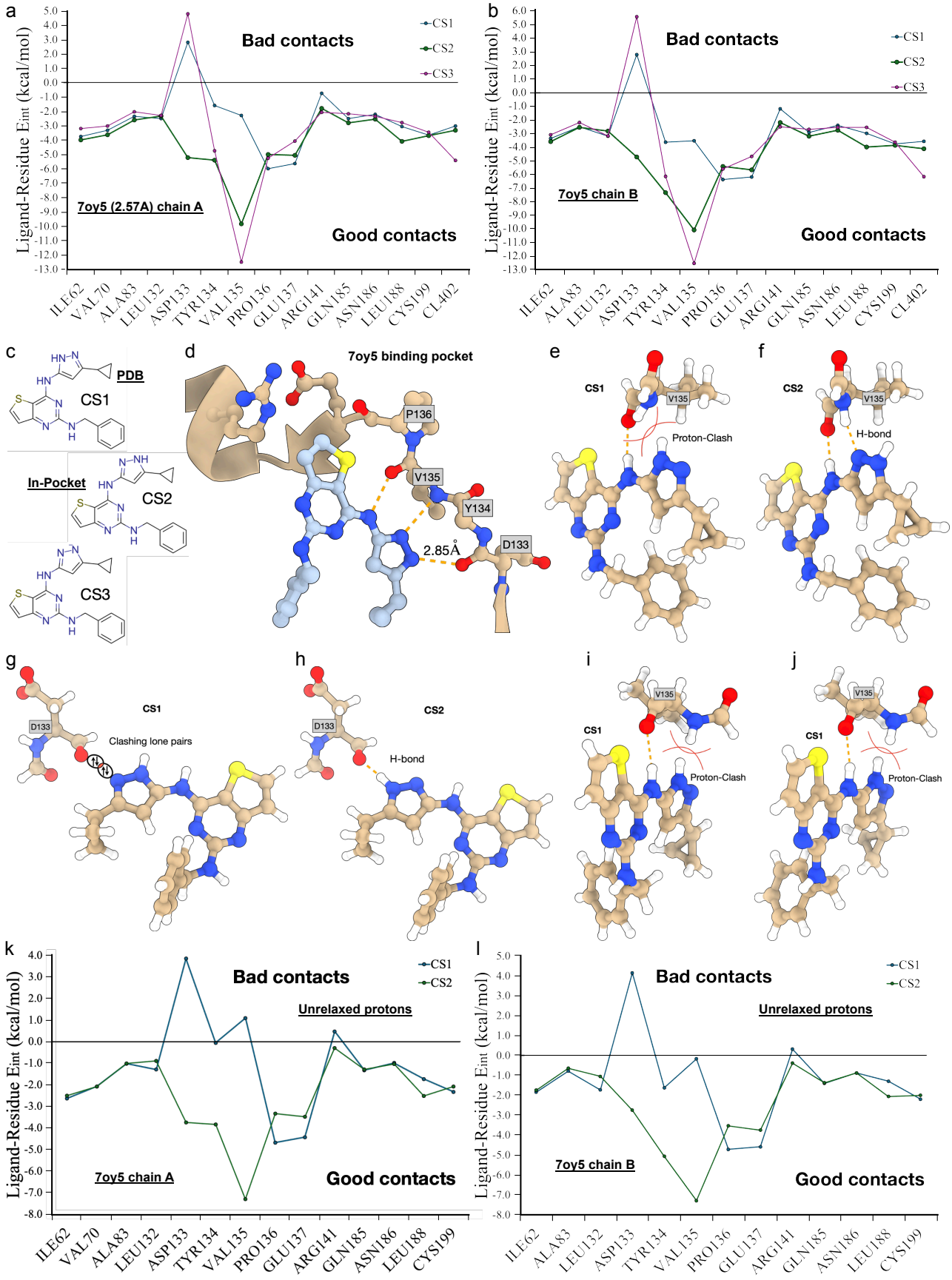


Figure S3: IPA data for a pyrazole-containing ligand bound to GSK3β, data from PDB-ID 7oy5.[34] a) The ligand-residue interaction energy plot for chain A shows that, of the 3 chemical states identified, only CS2 does not clash with the kinase's hinge. c) The same plot for chain B. c) The IPA-identified ligand chemical states. CS1, the PDB-deposited topology, differs from the IPA-identified chemical state, CS2. d) Pocket schematic representation with possible interactions from the different chemical states. e) The IPA-optimised CS1-V135 pair shows a proton clash with the main chain amide, avoided during the geometry optimisation by changing the hybridisation of the amide's nitrogen. f) The CS2-V135 pair shows the hydrogen bond missed by the PDB-deposited topology. g) The CS1-D133 pair shows a clash of lone pairs. Unlike a proton clash, the lone pairs cannot avoid one another. h) The CS2-D133 interaction shows the formation of a hydrogen bond. i) The chain A and j) chain B representations for the CS1-V135 optimised pair with all protons included in the bias potential. The protons cannot avoid each other in this situation, and the clash is more prominent. The ligand-residue interaction energy plot for k) chain A, and l) chain B, when including protons in the bias potential. Here, only minor refinements from the initial structures are possible, and the clashes are more evident.

S4 – Isoquinolone-lactam equilibria in the PDB

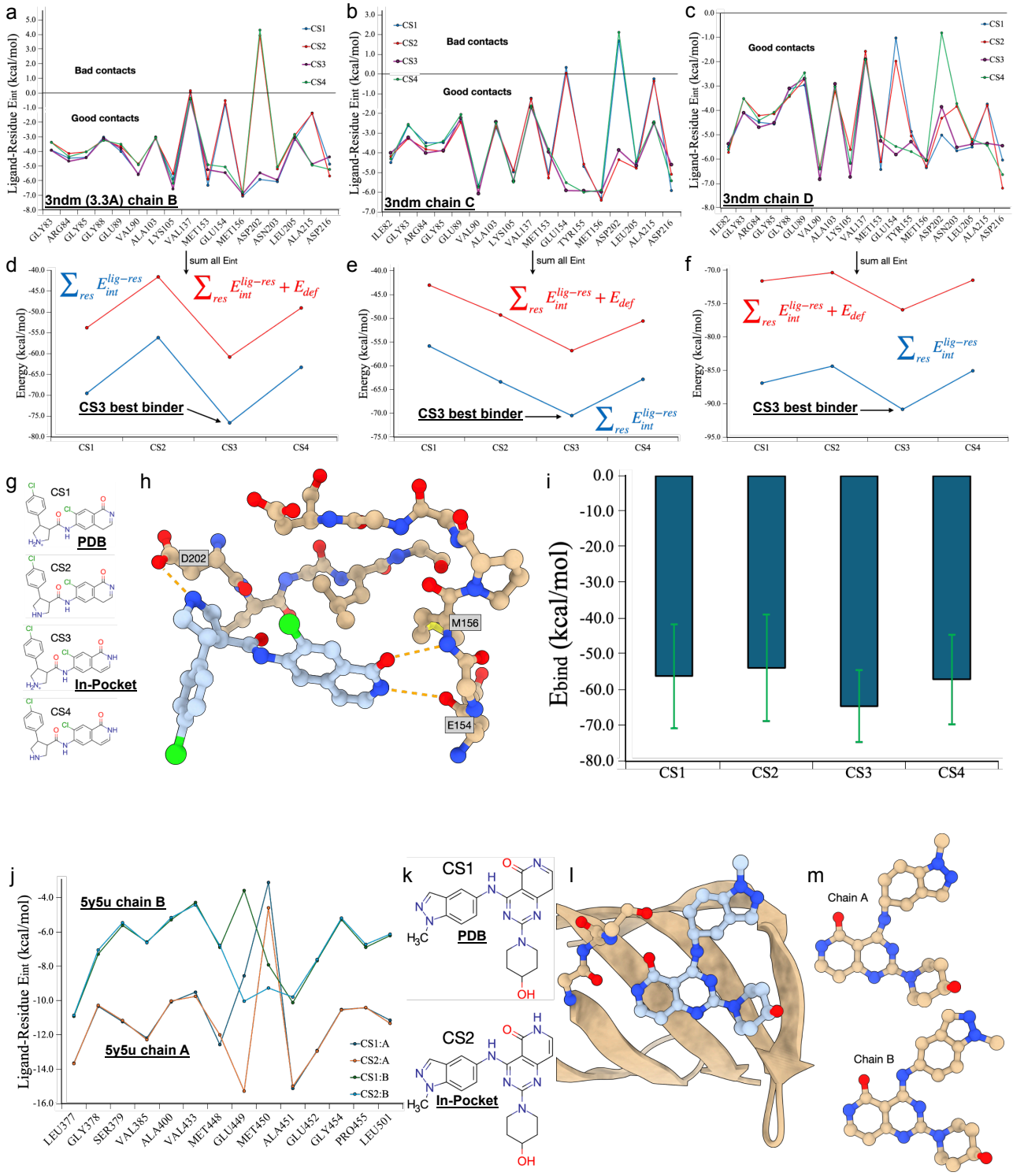


Figure S4: IPA data for isoquinolone-lactam equilibria, data from PDB-IDs 3ndm (a-i)[35] and 5y5u (j-m).[36] The ligand-residue interaction energy plot for chains a) B, b) C, and c) D of the 3ndm model. Note that no ligand was identified in chain A. The summed interaction energies (blue) and the IPA binding energies (red) for chains d) B, e) C, and f) D show that the lactam CS3 offers a better representation of the ligand in the pocket of the ROCK1 kinase. This contrasts with the PDB-deposited topology CS1. g) The IPA-identified ligand chemical states. h) ROCK1's pocket representation with the 3ND ligand. i) The statistical analysis of the IPA binding energies shows that, within the variability of one PDB

file (3 chains), CS3 is clearly the best binding conformational mode. j) The ligand-residue energy plots for another isoquinolone-lactam equilibrium, 5y5u. Both chains are represented in one plot. There is a remarkable difference in the range of stabilising interactions for each chain. k) The ligand chemical states identified by IPA. l) Tyrosine kinase's pocket representation with the chain A ligand. A significant distortion in the central linker is observed. m) Comparing the ligand structures in the two chains of the PDB 5y5u shows that chain A is unreliable. Analysing all the data, it is clear that although the isoquinolone CS1 does not induce any clash, it lacks hydrogen bond interactions with E449 and M450.

S5 – The protonation states of sulfonyl-ureas

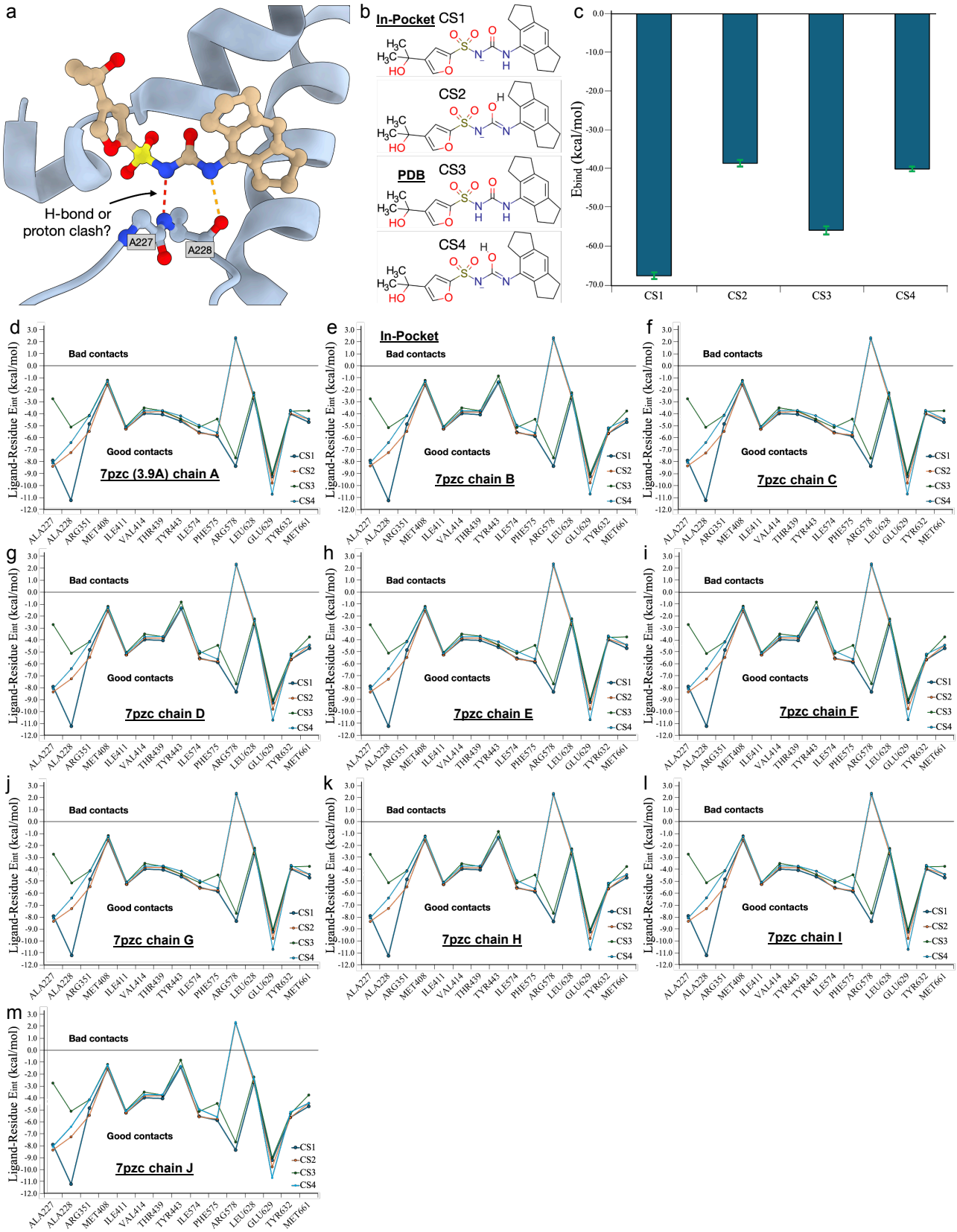


Figure S5: IPA data for the sulfonyl-urea ligand in the PDB 7pzc.[37] a) The ligand-in-the-pocket representation shows a potential clash between one of the sulfonyl-urea protons and the main chain nitrogen of A228. b) The IPA-identified ligand chemical states. c) IPA binding energy analysis over the 10 chains available in the PDB file. The calculations

systematically identify CS1 as the most likely chemical state of the sulfonyl-urea ligand bound to the protein. d-m) The ligand-residue interaction energy plots for each chain. Though proton abstraction from the sulfonyl-urea might initially seem odd, this result greatly agrees with the sulfonyl-urea pK_a in the range of 4-6.[38]

S6 – Protonating amines and flipping pyridines

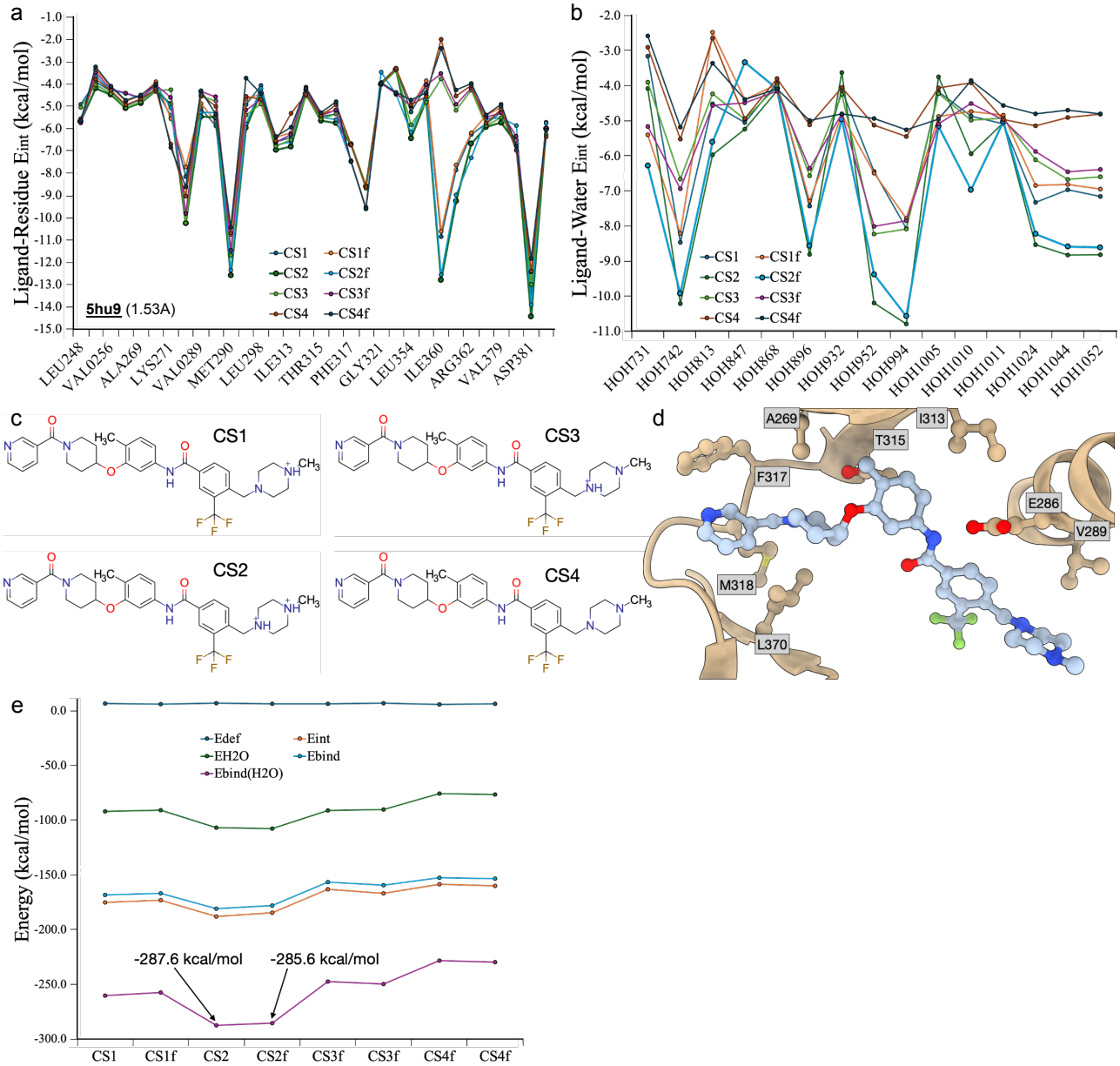


Figure S6: IPA analysis of an N-methylpiperazine ligand with a potential pyridyl flip. a) The ligand-residue interaction energy plot for the protein-ligand complex in the PDB 5hu9.[39] b) The ligand-water interaction energy plot for the structural model in 5hu9. The analysis was performed for the four IPA-identified chemical states with and without terminal pyridyl flip (suffix "f"). The ligand-residue interaction pattern is similar for all chemical states except around I360. This indicates an advantage for protonating the most external nitrogen in the piperazine ring. Conversely, the ligand-water interactions seem stronger for CS2 for both pyridyl ring orientations. c) The IPA-identified ligand chemical states. d) The pocket model for the complex in 5hu9. e) The summed energy analysis for this complex: The deformation energy (E_{def}), the interaction energy with solvation waters (E_{H2O}), the summed ligand-residue interaction energies (E_{int}), the IPA binding energy ($E_{bind} = E_{int} + E_{def}$) and the IPA binding energy with the water contributions ($E_{bind(H2O)} = E_{bind} + E_{H2O}$). The last quantity differs by 2.0 kcal/mol for the two flips of CS2, indicating that it is likely that both pyridyl ring orientations are biologically stabilised and observed.

S7 – Curcumin chemical states identified by IPA

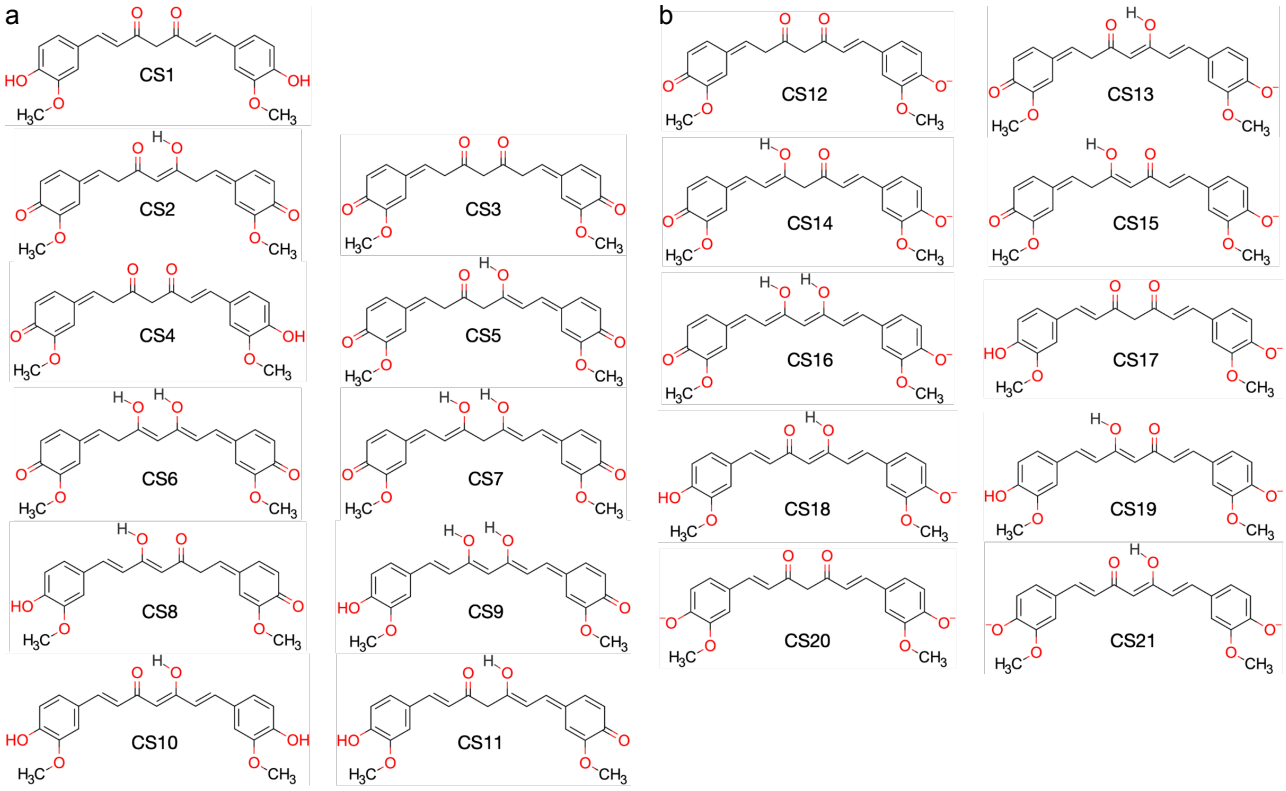


Figure S7: The manyfold curcumin chemical states identified by IPA. a) The neutral chemical states. b) Charged chemical states.

S8 – Curcumin hydrogen bond network optimisation using IPA

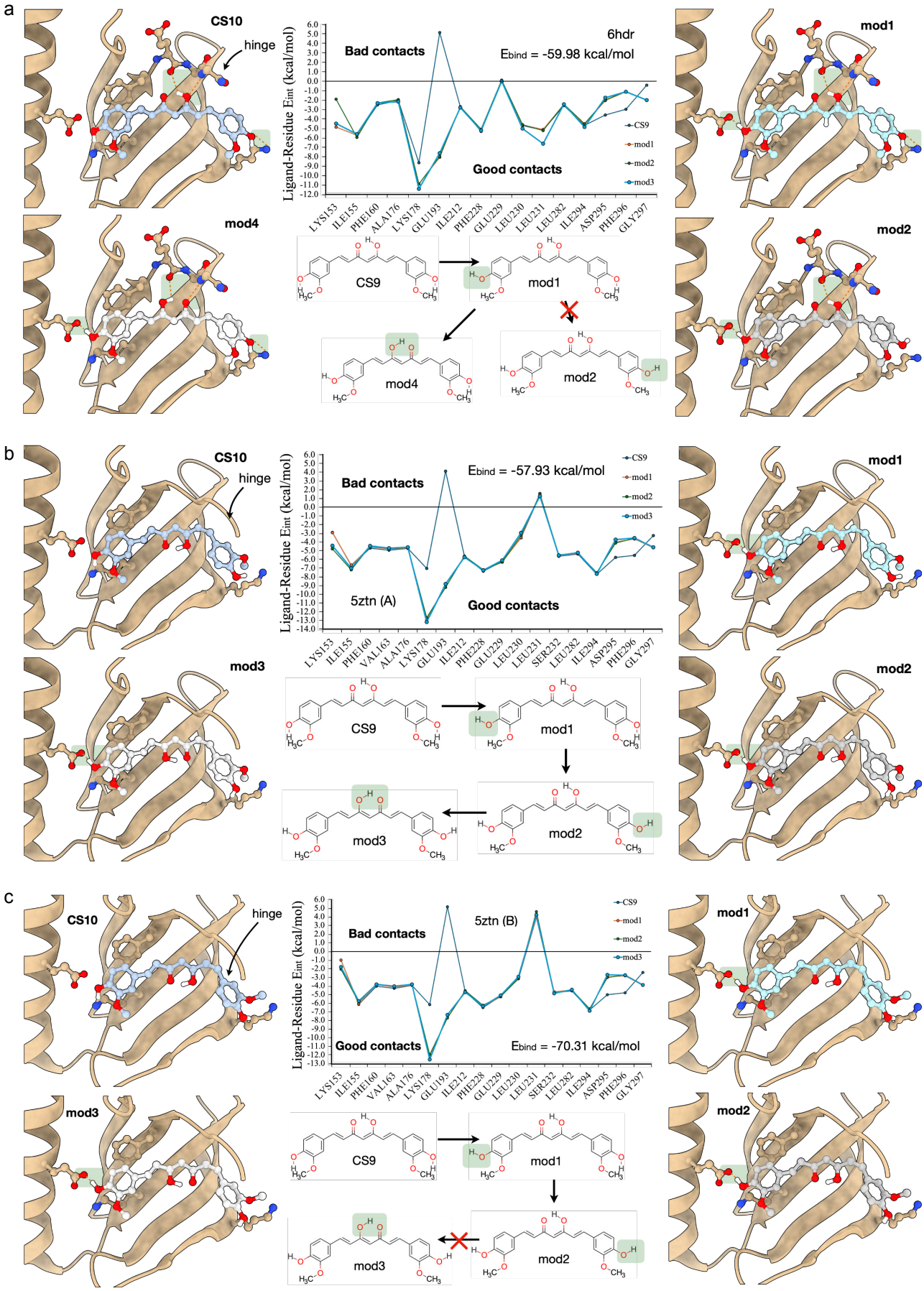
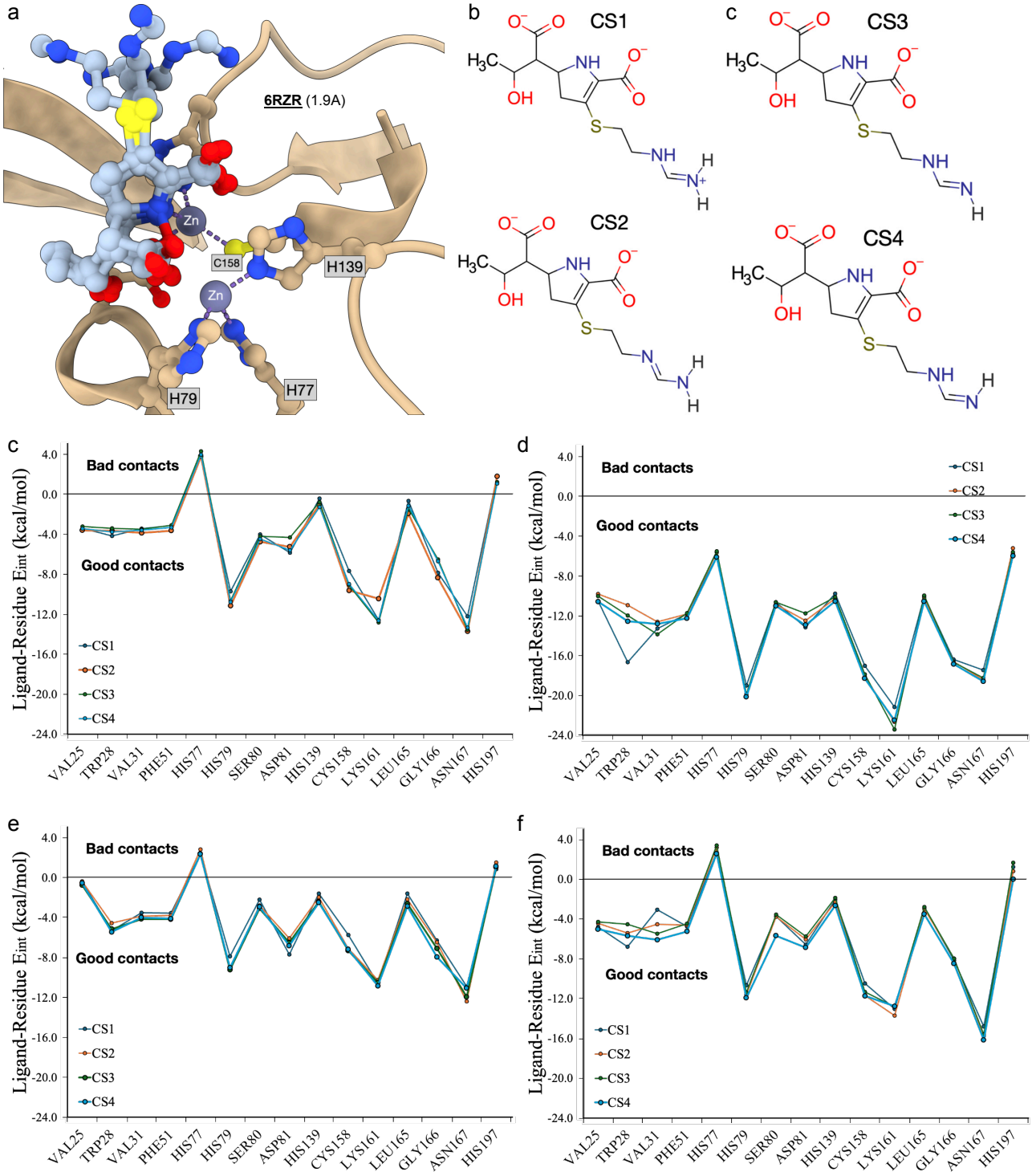


Figure S8: IPA of curcumin with different protonation patterns consistent with curcumin's CS10 chemical state. The impact of orientation and proton jumps on the protein-ligand complex stability can be followed by the ligand-residue interaction energy plots in the centre of each block. Structural models are placed to the left and right of the interaction energy plots. This data supplements the information in Figure 4 of the main manuscript. a) The hydrogen bond network optimisation for the structural model 6hdr. b) The hydrogen bond network optimisation for the structural model 5ztn, chain A. c) The hydrogen bond network optimisation for the structural model 5ztn, chain B.

S9 – Analysing the chemical and conformational dynamics of a metal-containing binding pocket: inhibiting a β-lactamase



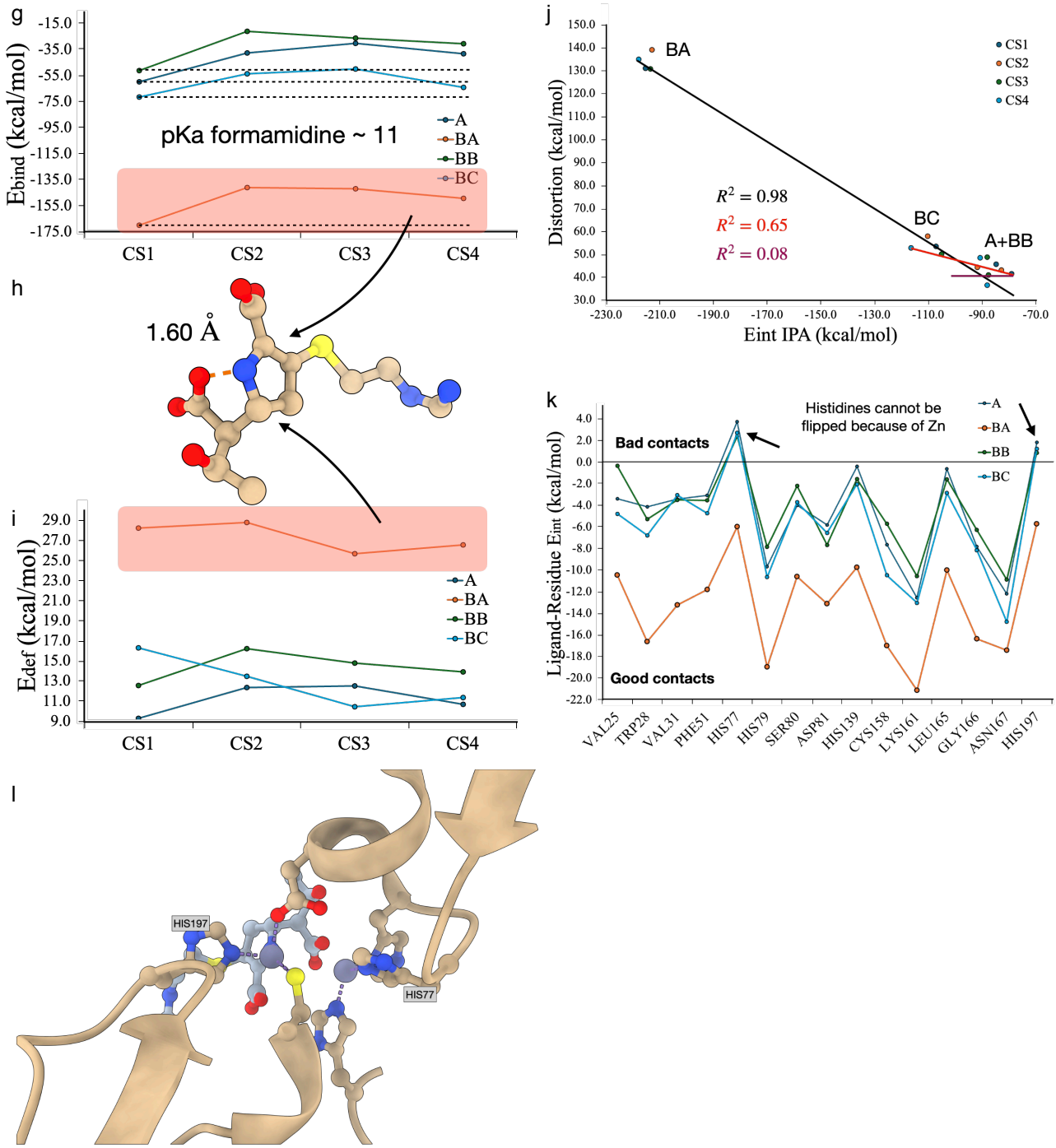


Figure S9: Analysing the alternate locations of a ligand next to a metal-containing binding pocket, PDB-ID 6rzr.[18] a) The pocket representation, with an overlay of the four structural models: chain A and chain B, the latter with three alternate binding poses (BA, BB, BC). b) The four IPA-identified ligand chemical states available. Note that we purposely excluded the carboxylic acid protonation states. The ligand-residue interaction plots for c) chain A and the chain B models d) BA, e) BB, and f) BC. Although the ligand-residue interaction energy plot for the BA model shows a pattern similar to other models, this is shifted towards lower energies, indicating something odd with this model. g) IPA binding energies also show a different behaviour for the BA model. h) Ligand-only representation of the BA model, evidencing an N-O short contact that is almost within the distance of a covalent bond. This inaccuracy in the structural model raises the i) ligand deformation energies for all chemical states identified. j) The distortion Vs. the IPA-summed ligand-residue interaction energies plot shows a good correlation when the data referring to the structural model BA is included. The

correlation is negatively impacted by excluding model BA's data, but in this case, there is still a strong Pearson correlation between model distortion and the summed IPA interaction energies. This indicates that one of the models is still contaminated with low experimental data resolution. Excluding the BC-model dataset destroys any correlation between distortion and the summed interaction energies, showing that models A and BB are the most reasonable choices for further studies. k) The ligand-residue interaction energy plots for CS1 in several structural models. This shows that, of the two remaining models, chain A is the QM most indicated pocket representation of this complex. IPA shows two poor contacts associated with two histidine residues. l) Analysing the structural model shows that the histidines cannot be flipped; otherwise, the interactions with the zinc ions would be impacted.

S10 – Venetoclax binding to BCL-2. Conformational freedom or crystal artefact?

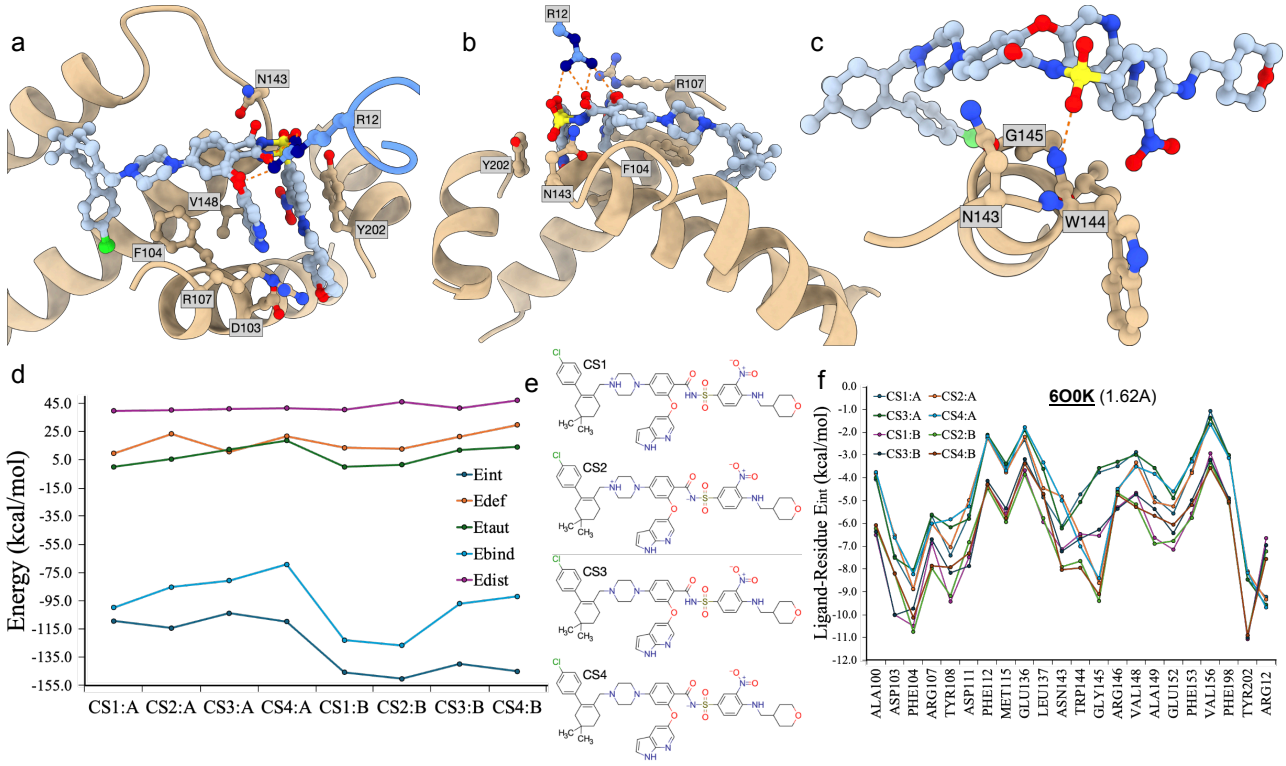


Figure S10: Two venetoclax models in BCL-2, PDB-ID 6o0k.[19] a) The pocket of venetoclax in BCL-2 has two alternate ligand representations (A and B). Residue R12 is included from a symmetry copy in the crystal. b) Alternative view/representation of the binding pocket, showing the interaction pattern between venetoclax and the residue R12. c) The interaction of venetoclax with G145. By deprotonating the sulfonyl-amide, charge density is increased in the sulfonyl group, strengthening the hydrogen bond to G145. d) Global IPA metrics for the two structure models. This plot shows that all distortion energies are similar and within the same magnitude. This rules out distortion as the driver for the shift of the interaction energy plot in f), and it leads to the conclusion that alternate B offers a better structural model of venetoclax in the pocket of BCL-2. This is also confirmed by the IPA binding energy, which contains the summed ligand-residue interaction energies, the ligand deformation, tautomerisation, and isomerisation energies. Of the four chemical states investigated, those with the ammonium group are particularly favoured (vs. neutral amine). Deprotonation of the sulfonyl-amide group leads to an additional stabilisation by 4.0kcal/mol, by strengthening the hydrogen bond with G145's main chain nitrogen. Additional stabilisation is obtained by interactions with water molecules, see supplemental tables and plots in the supplements folder. We stress that one of the main differences between models A and B is the interaction of the ligand with R12. As the target BCL-2 is not a biological dimer, and the interactions with R12 result from symmetry copies, it is likely that this is a crystallographic artefact and that the experimentally observed alternate conformation A is not biologically relevant. e) The four venetoclax chemical states we considered in the analysis. f) The ligand-residue interaction energy plots show a potential shift in the interaction energy pattern of the two alternate models we are analysing, alternate A and alternate B. The almost indistinguishable distortion energies between the two models (see above) show this is not caused by ligand distortion or model quality, but rather differences in the binding pose.

S11 – Conformer states of 3zy2.

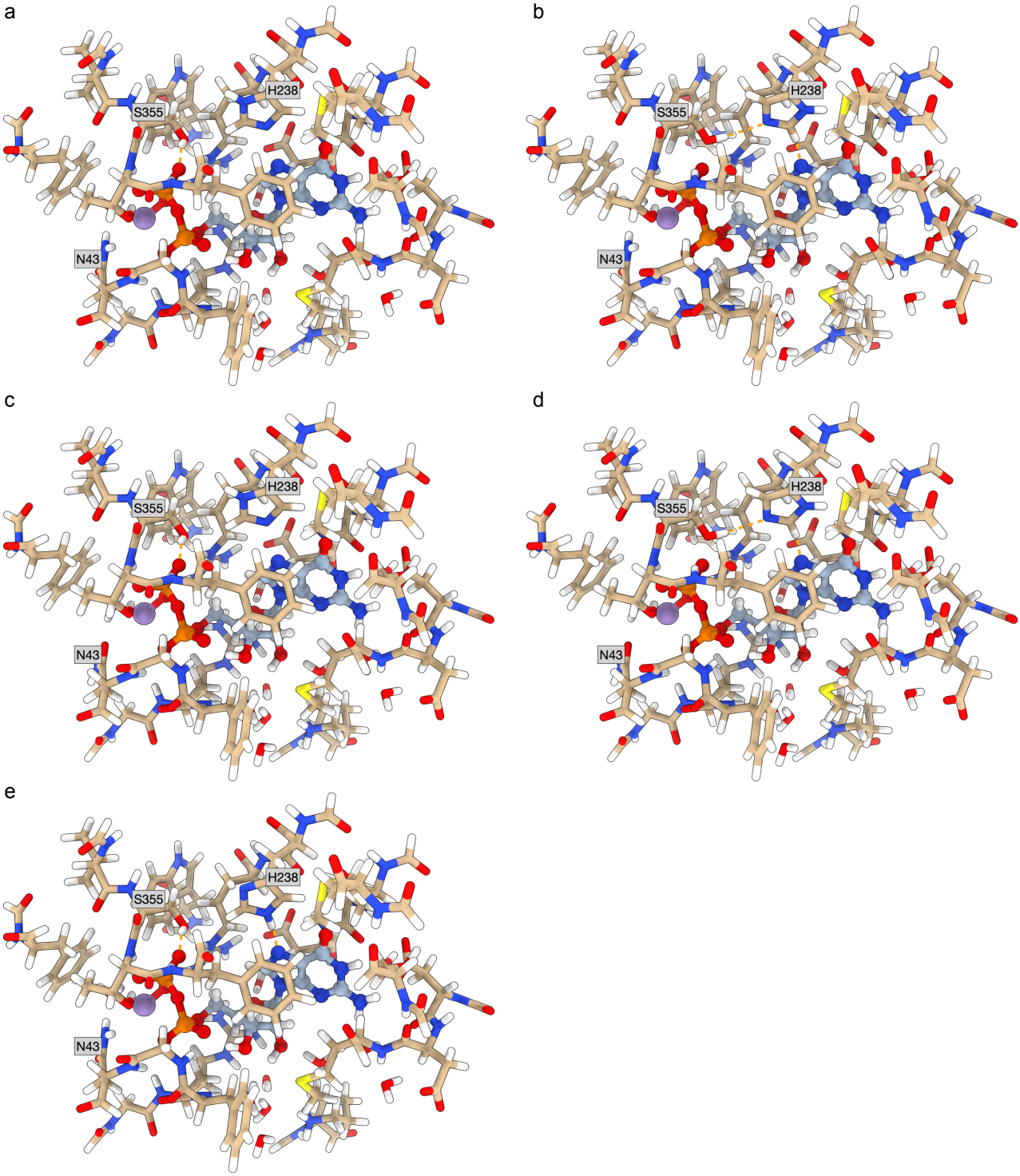


Figure S11: Structural representation of the five 3zy2 conformer states tested. a-e) represent, respectively, the states 1-5 referenced in **Figure 5i** and **5j** obtained by combinatorially flipping H238 and N43. The last state changes the protonation state of His238 from δ to ε.

S12 – IPA analysis of transition metal-containing systems.

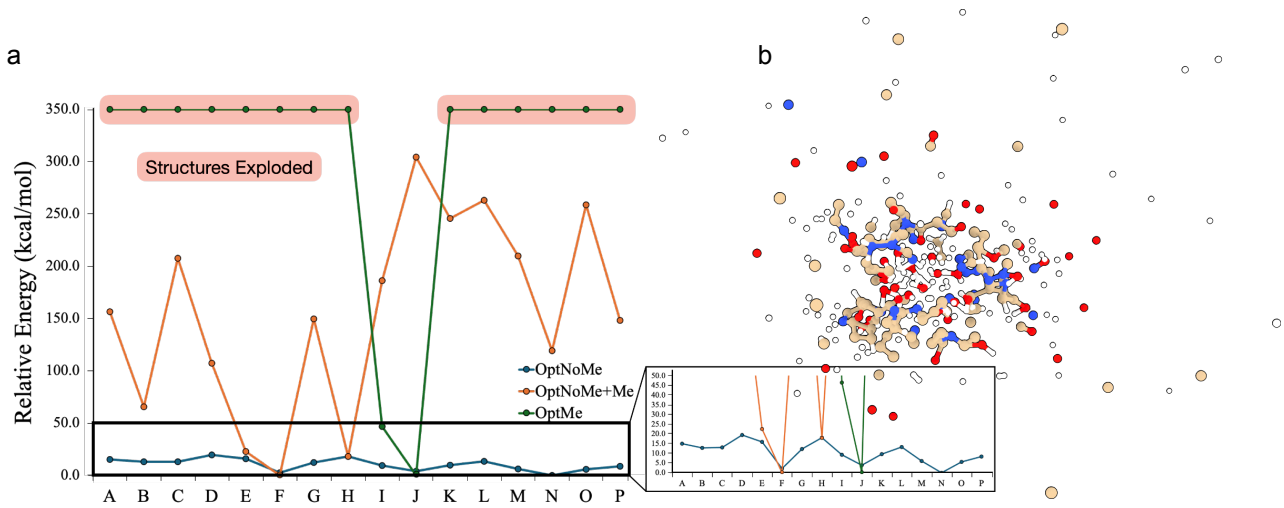


Figure S12: Retrieving chemically meaningful information for large atomic systems. a) The energy plot of all conformational states considered when analysing the PDB-ID 3wml. Three curves are shown: a blue curve, where the fragments were optimised without the metal centres. Though reasonably flat, the energy surface is dominated by a single conformer. Introducing the metal ions on the previously optimised structures, the orange curve severely destabilises several conformers, particularly the expected energy minimum. Finding the energy minimum requires optimising the protein fragments in the presence of metal ions, represented in the green curve. In the presence of metal ions, only two structures can be optimised. Despite the constraining potential used, most structures ended up exploding, for which we pragmatically assigned a relative energy of 350.0 kcal/mol. b) Example of one of the unstable conformer states probed that exploded, despite the bias potential used during the optimisation.

S13 – IPA analysis of 3wml.

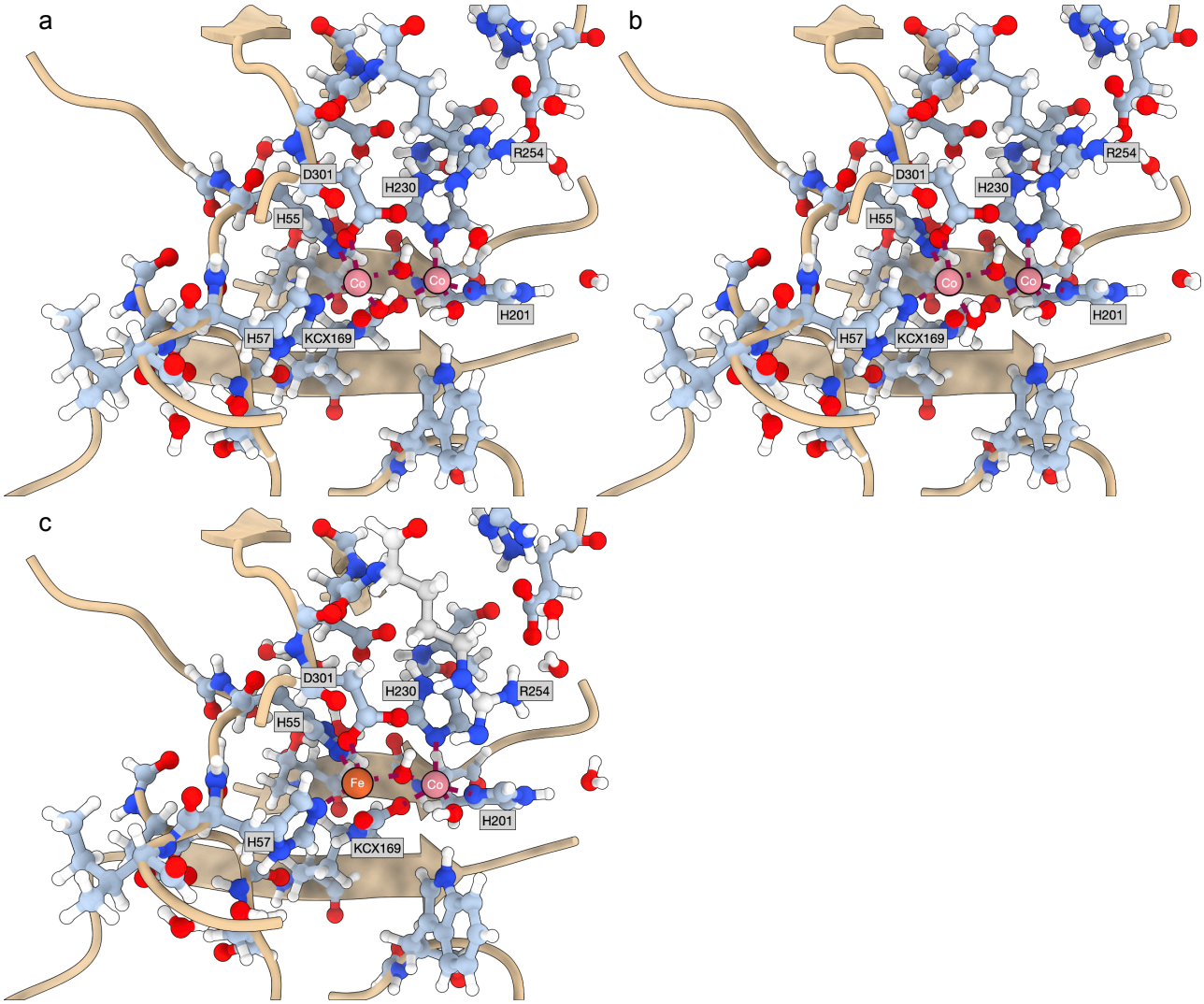


Figure S13: The three states of 3wml. Complete characterisation of the states in 3wml. a) The Co-Co state with a hydroxyl group coordinated to the leftmost cobalt ion. b) The Co-Co state with a coordinated water molecule. c) The Fe-Co state, with a change in conformation of the residue R254. This residue is furthermore marked with a different colour.

S14 – IPA analysis of 1wug, an NMR structure.

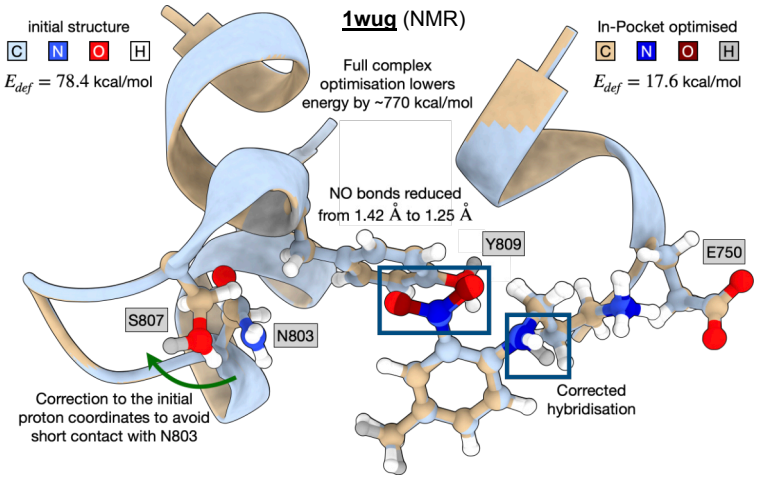


Figure S14: IPA refinement of the NMR structure 1wug. The initial ligand's deformation energy of 78.4 kcal/mol can be reduced to 17.6 kcal/mol using IPA, resulting from corrected N-O bond distances in a nitro-group and fixes in atomic hybridisations.

Structural distortions in nitro groups are pervasive in PDB-deposited models. In addition to bond length discrepancies, asymmetric N-O bond distances are frequently observed within one monomer, which are inconsistent with the electronic structure of nitro groups, even in the anisotropic environment of a protein's pocket. To further investigate these distortions, we analysed the 1wug protein-ligand complex (CCD-ID: NP1),[40] which features a nitro-containing ligand bound to a bromodomain (**Figure S14**). The IPA refinement significantly reduced N-O bond distortions while optimising the hydrogen-bonding network. These structural corrections translated into substantial energetic improvements: the overall system stabilised by 770.0 kcal/mol due to a more favourable network of interactions, while the ligand-specific deformation energy dropped by 60.8 kcal/mol. Notably, IPA also corrected the hybridisation states of the aniline-like nitrogen, a required adjustment since classical protonation algorithms rely exclusively on geometric heuristics. These refinements emphasise the importance of QM corrections in ensuring chemically accurate ligand representations within protein-ligand complexes.

S15 – IPA analysis of large molecular assemblies.

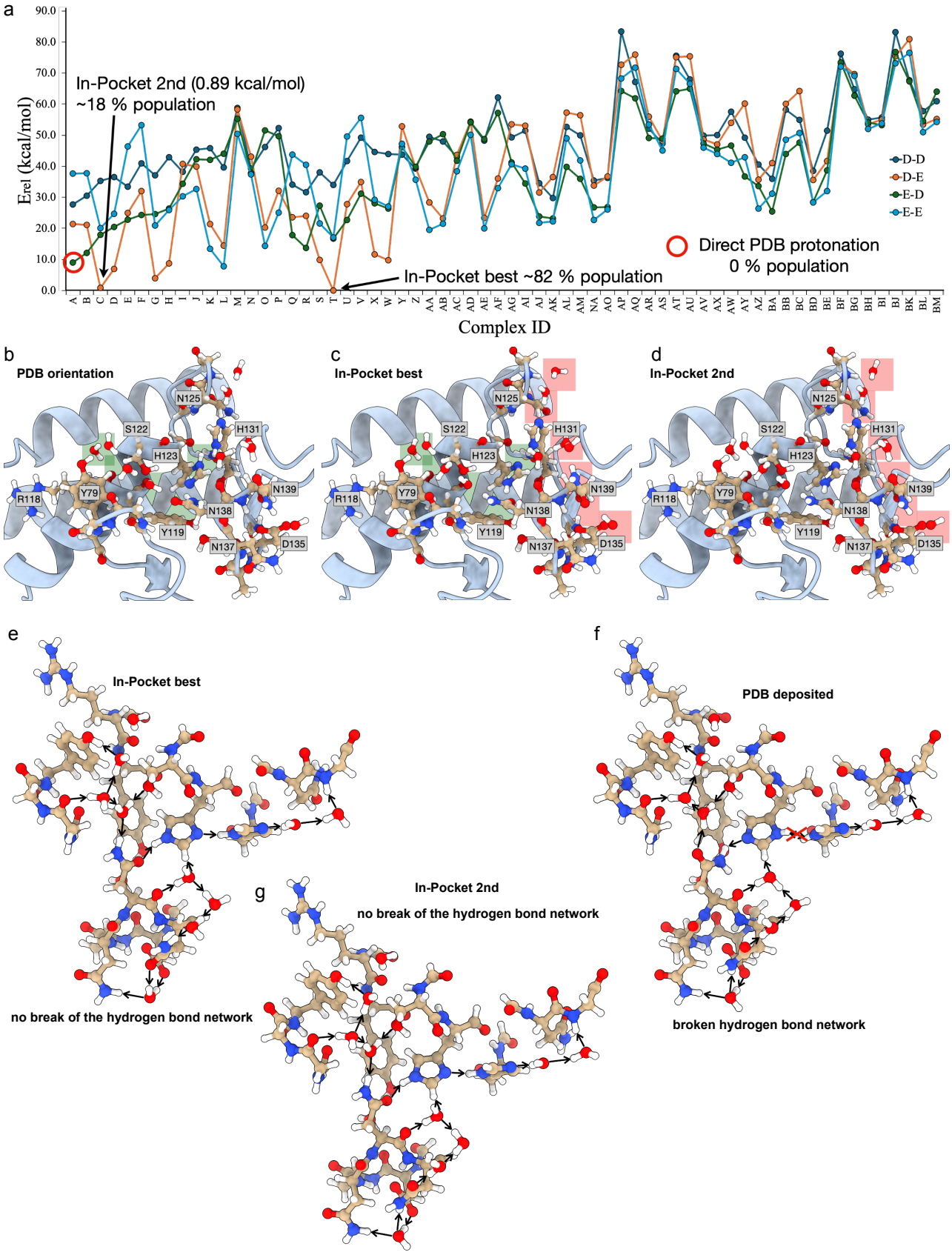


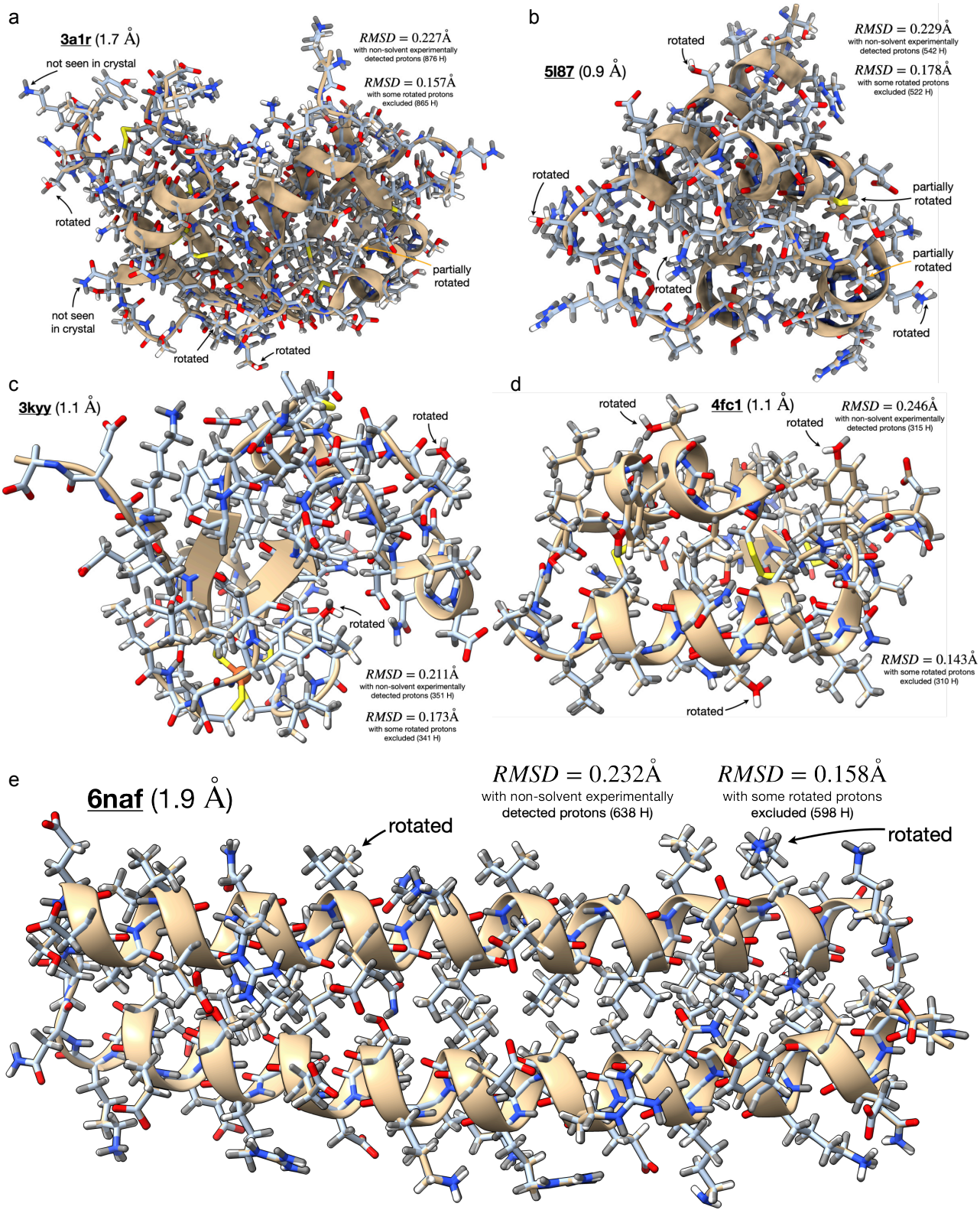
Figure S15: Alternative hydrogen-bond networks in the PDB 1cem. a) The relative energy of each network state identified for the selected fragment of the PDB structure 1cem. For each histidine protonation pattern (δδ = D-D; δε = D-

E; εδ = E-D; εε = EE), 64 orientational conformers were identified and refined with IPA. Depending on the protonation pattern selected for the histidine residues, the PDB-deposited structure is at least 10.0 kcal/mol higher in energy than the lowest energy state, indicating that this structure is thermodynamically irrelevant to describe the system. A direct PDB protonation corresponds to proton addition according to the PDB-deposited topology, following a δ-protonation of histidines. IPA's fragment mode improves b) the initial PDB-deposited conformer. The green shaded areas locate the major differences recorded between the hydrogen-bond refined PDB structure and the c) IPA minimum energy structure. d) The second-lowest energy system identified by IPA. The red-shaded areas show differences with respect to the IPA energy minimum. The second most stable structure is about 0.89 kcal/mol less favourable than the minimum found. e) The hydrogen-bond network of the lowest energy chemical state identified for the 1cem fragment we analysed. An uninterrupted network of 18 hydrogen bonds justifies the stability of this conformer. f) The hydrogen-bond network for the PDB-deposited chemical state shows an interruption due to a proton clash between the histidine residues. This structure is thermodynamically irrelevant to describe the system. g) The second most stable IPA identified state also shows an uninterrupted network of 18 hydrogen bonds, though slightly less favourable than the energy minimum.

To assess IPA's ability to analyse extensive hydrogen-bond networks in biological systems, we examined 256 proton-network orientations in the PDB model 1cem[41] (**Figure S15**): residues H123 and H131 may be δδ, δε, εδ, or εε protonated, yielding 64 possible structures obtained by combinatorially flipping His, Asn, and Gln residues. Inherently, the surrounding water network adapts to side chain relative orientations. Of all conformations generated, only structures A–δδ, δε, εδ, and εε–are consistent with the deposited PDB structure.

Our calculations reveal that the PDB structure with δδ His protonation is nearly 30.0 kcal/mol higher in energy than the lowest-energy conformer (**Figure S15a**). Boltzmann statistics indicate that a naïve, direct protonation of the structure according to the PDB topology is thermodynamically irrelevant. The elevated energy results from a proton clash that disrupts the hydrogen-bond network (**Figures S15b** and **S13b**). An εδ His protonation moderately improves the system's energetics, yet it remains 10.0 kcal/mol above the lowest-energy state and is similarly biologically irrelevant. The most stable conformation was identified after three Asn flips (N125, N138, and N139) and adopting a δε His protonation (T conformer, **Figures S15a,c**, and **S15e**). This state accounts for approximately 82% of the conformer population. The second most populated conformer, featuring a single N139 flip from the PDB structure with δε His protonation (**Figures S15a,d**, and **S15g**), lies 0.9kcal/mol above the minimum energy state and represents ~18% of the total population. To benchmark further, we used MolProbity[42] and WHAT_CHECK[43]and observed agreement with the IPA results.

S16 – Additional IPA validation of fully optimised proton networks.



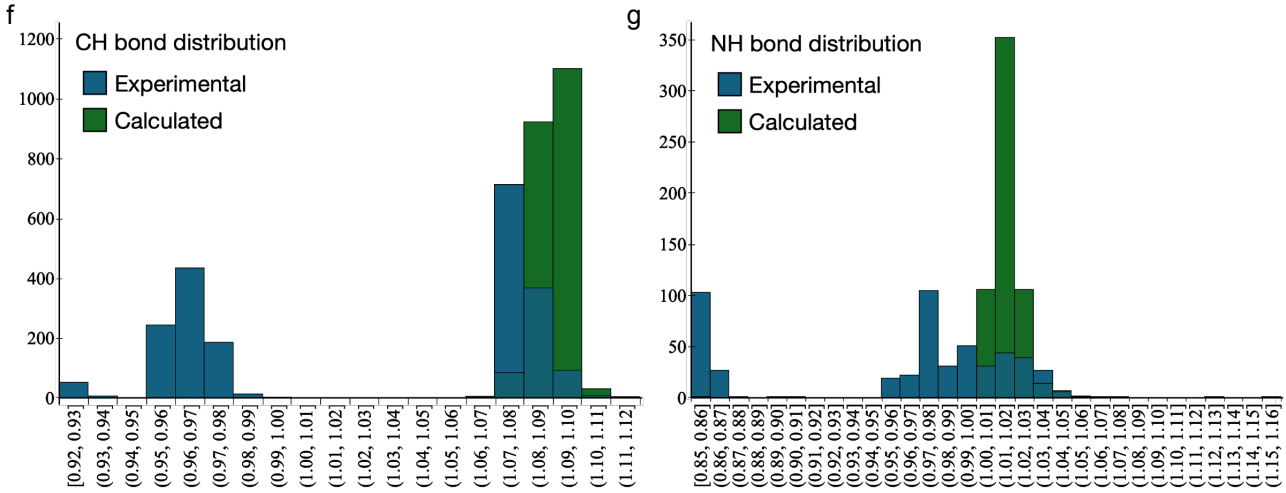


Figure S16: Additional IPA validation through proton-refined coordinates. Validation using the PDBs a) 3a1r,[44] b) 5l87,[45] c) 3kyy,[46] d) 4fc1,[47] and e) 6naf.[48] These cases compare the IPA-optimised proton network against experimentally declared proton coordinates. Besides a few rotamers observed on the protein's solvent-exposed surface, the two sets of proton coordinates are in great agreement. The structures, spanning various resolutions, were re-protonated to introduce randomness in initial proton placements, while all non-hydrogen atoms were preserved to best reflect the typically experimentally observed system. The accuracy of IPA-optimised structures was assessed by computing the root-mean-square deviation (RMSD) between initial and final proton coordinates, yielding values of 0.21–0.25Å. This analysis included only protein and, where present, small-molecule proton coordinates, excluding solvent and crystallisation additives. A closer inspection reveals that major deviations arise primarily from freely rotating groups, such as terminal methyl or ammonium moieties. Additionally, some solvent-exposed hydroxyl groups (*e.g.*, tyrosine) remained flipped by 180º due to their randomised starting orientations. Correcting these orientations to match experimental data would require overcoming an activation barrier, which is unfeasible for geometry optimisation algorithms. Excluding these protons from the RMSD calculation reduced the experimental-calculated RMSDs to 0.14-0.18Å, highlighting the precision achievable with IPA. f) Comparing the IPA-optimised CH bond distance distributions against the experimentally observed ones. The former shows a unimodal distribution, whereas the latter is multimodal. g) Comparing the IPA-optimised NH bond distance distributions against the experimentally observed ones. Again, the former shows a unimodal distribution, whereas the latter is multimodal.

S17 – The IPA Algorithms.

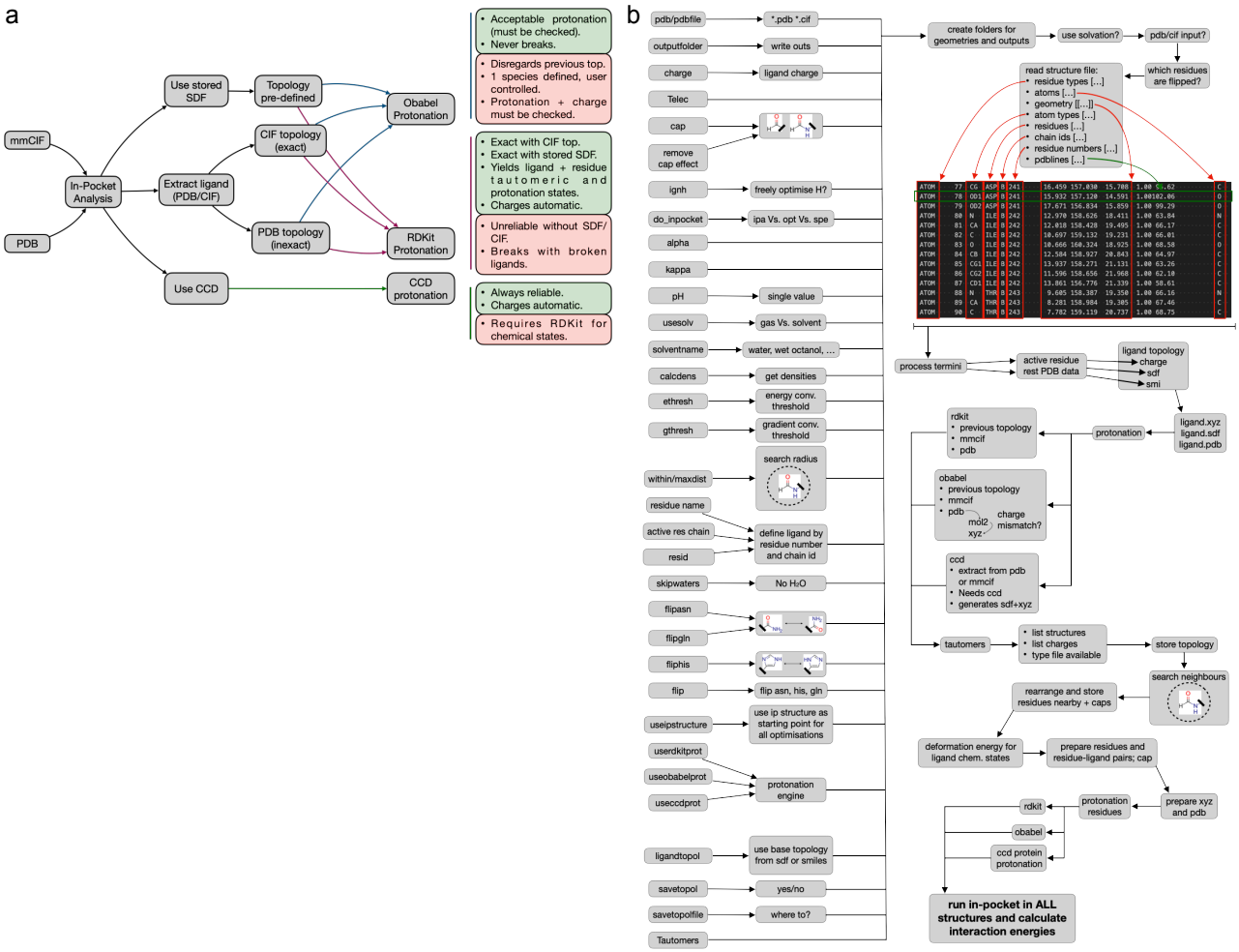


Figure S17: Schematic Representations of the IPA algorithms and an overview of IPA options. a) Schematic representation of how the IPA algorithm works. It is possible to start the process using a PDB or a mmCIF file. Different algorithms may be used for protonating structures and to proceed with further processing. Each algorithm's advantages (green) and disadvantages (red) are indicated next to it. b) Schematic overview of how IPA works and an overview of all options available.