CosolvKit: a versatile tool for cosolvent MD preparation and analysis

Niccolo' Bruciaferri[1], Jerome Eberhardt[3,2,1], Manuel A. Llanos[1], Johannes R. Loeffler[1], Matthew Holcomb[1], Monica L. Fernandez-Quintero[1], Diogo Santos-Martins[1], Andrew B. Ward[1], Stefano Forli[1]

Abstract

Cosolvent molecular dynamics (MD) are an increasingly popular form of simulations where small molecule cosolvents are added to water-solvated protein systems. These simulations can perform diverse target characterization tasks, including cryptic and allosteric pocket identification and pharmacophore profiling, and supplement suites of enhanced sampling methods to explore protein conformational landscapes. The behavior of these systems is tied to the cosolvents used, so the ability to define diverse and complex mixtures is critical in dictating the outcome of the simulations. However, existing methods for preparing cosolvent simulations only support a limited number of predefined cosolvents and concentrations. Here we present CosolvKit, a tool for the preparation and analysis of systems composed of user-defined cosolvents and concentrations. This tool is modular and agnostic of the MD engine and force field used, offering access to a variety of generalizable small molecule force fields. To the best of our knowledge, CosolvKit represents the first generalized approach for the construction of these simulations.

[1] Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, 92037 California, United States;

[2] SIB Swiss Institute of Bioinformatics, University of Basel, Basel, Switzerland (Present affiliation)

[3] Biozentrum, University of Basel, Basel, Switzerland

Intro

Molecular dynamics methods are a valuable tool for modeling proteins to identify and characterize ligandable sites during a structure-based drug design campaign, or characterize protein-protein interaction surfaces. In the absence of the relevant binding partner, however, the relevant conformations, such as those that would expose a cryptic pocket or otherwise organize holo-like conformations, may be high energy and require prohibitively long simulations to sample. A variety of molecular dynamics methods have emerged to enable enhanced sampling of these states[1,2], but often significant expertise is required to perform and analyze them, and lack the ability to profile specific interactions with the binding partner.

Cosolvent simulations are a simple yet powerful alternative, in which additional small molecule probes besides water are introduced into the system. The addition of cosolvent compounds to the simulation can enhance protein motions, especially those related to the identification of cryptic pockets[1,3,4]. These cosolvents can also be treated as probes, identifying ligandable sites and associated fragments, or standing in for pharmacophores to profile a site. This has led to the development of tools to analyze the resulting trajectories, such as Cosolvent Analysis Toolkit[5] and Probeview[6], as well as a variety of methods to prepare these simulations.

However, existing methods, such as MixMD[6], MDMix[7], and SILCS[8], to prepare cosolvent simulations are limited in the scope of cosolvents and concentrations they can accommodate, to the point of having disjoint sets of cosolvents accessible between methods[9]. This stems from the need for precomputed forcefield parameters for the cosolvent. However, different cosolvent mixtures may be desirable for different tasks (e.g. where specific chemical species or interactions are of interest), and even for a well defined task such as cryptic pocket identification the ideal mixture may vary depending on the target[10]. Yanagisawa et al.[9] were motivated by this to attempt to identify and provide parameters for a universal cosolvent mixture, but had to contend with goals of generality of the cosolvent set and complexity of the resulting simulation. Additionally, while a universal cosolvent set would be beneficial to general exploratory simulations, there might be either arbitrary cosolvents of interest or atypical targets where user-defined cosolvents or concentrations (or both) may be critical to answering specific scientific questions.

A solution to these constraints on cosolvent identity and concentration would be to allow on-the-fly parameterization of cosolvents, and to provide tools to characterize the equilibration of the cosolvent mixture as an alternative to pre-equilibrated patches. Here, we describe CosolvKit, a toolkit for the preparation and analysis of cosolvent simulations built on OpenMM [11], and employing generalized small molecule force fields to parametrize the molecular probes (i.e. Espaloma[12], GAFF[13], and OpenFF[14]). While CosolvKit provides native support to run the simulations within OpenMM, it can also generate input files for other widely used MD engines, such as GROMACS[15], Amber[16], or CHARMM[17]. We further characterized the performance of CosolvKit simulations across diverse tasks, including cryptic pocket identification, protein-protein interaction profiling and mapping of designable active sites, and demonstrated how the provided diagnostics can aid in troubleshooting these complex simulations.

Results and Discussion

CosolvKit core features

To the best of our knowledge, currently available methods to set up cosolvent simulations rely on pre-equilibrated patches containing mixtures of water and cosolvent molecules. These patches only cover a limited variety of cosolvents and the concentrations are fixed at arbitrary values. The aforementioned limitations motivated the present work, which aimed to develop a comprehensive, flexible, reproducible, and easy-to-use solution to set up cosolvent simulations.

CosolvKit was developed in Python 3 leveraging the OpenMM framework[11] and provides ready-to-use scripts with a user-friendly command line interface to facilitate the creation and analysis of cosolvent simulations. At the same time, it also exposes a powerful API for full customization and integration in more advanced simulation pipelines.
The command line interface accepts a JSON configuration file, which enables users to define reusable simulation recipes to customize and parametrize cosolvent systems.

The pipeline to build a system consists of three main steps (Figure 1). It starts with the creation of an empty simulation cubic box containing the macromolecule of interest (i.e., protein or nucleic acid). While most cosolvent simulations are performed in the presence of a macromolecule, CosolvKit can define simulation boxes that do not contain any macromolecules, allowing to simulate fluids of arbitrary composition. If present, biological macromolecules can be sanitized by CosolvKit with the use of the PDBFixer application[11]. Optionally, CosolvKit can add membranes around the receptor through OpenMM[11], either by using the default lipids provided or by supplying a pre-equilibrated patch of the desired bilayer composition. The membrane is built along the z-axis, requiring macromolecules to be already aligned with respect to the hydrocarbon core of the lipid bilayer (e.g. as for structures retrieved from the Orientation of Proteins in Membranes database[18]).
CosolvKit is intended to be as general and engine-agnostic as possible. Therefore, when creating a cosolvent system, the user can specify the output format for topology and position files. All major MD engines are currently supported, such as Amber[16], GROMACS[15], CHARMM[17], and OpenMM[11].
One of the key features of CosolvKit is the flexibility to parametrize cosolvent molecules using different force fields for small molecules, such as GAFF[13], Espaloma[12], or OpenFF[14]. Both cosolvent types and force fields can be defined as JSON recipe files and supplied to the command line, facilitating the setup process and its reproducibility. Then, 3D coordinates for cosolvent molecules are created from SMILES strings using the RDKit python library and placed within the box, maintaining a buffer distance of 3.5 Å from the macromolecule, if present. The user specifies the desired concentration either in Molar units or by the absolute number of cosolvent molecules. When a

concentration is provided, the number of cosolvent copies to place is computed based on the box accessible volume, (excluding the volume of the macromolecule if present) and the concentration requested. To avoid a lattice-like distribution and ensure a random uniform placement of cosolvent molecules a Halton sequence[19] generator is used to define their position in 3D space, and a random rotation is applied to each cosolvent molecule. A minimum distance of 2.5 Å between any cosolvent molecules is enforced.

The last step is the system solvation. The default solvent used is water, but the user can specify any solvent by providing its SMILES strings.

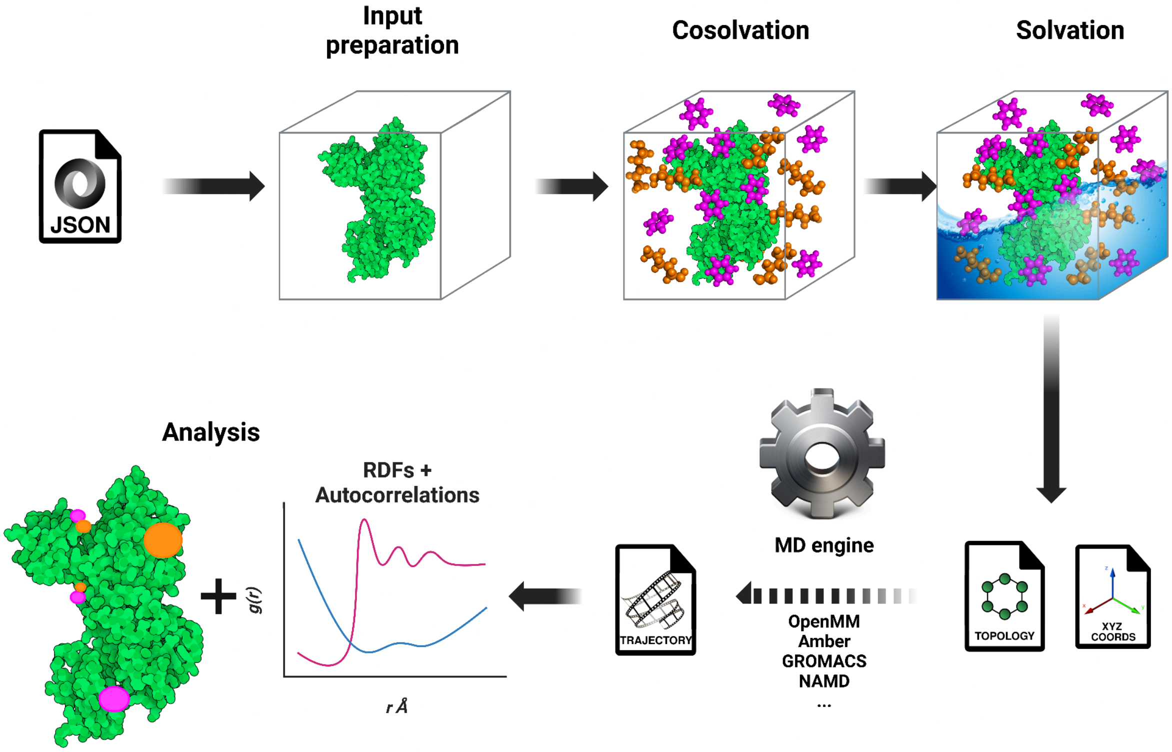

Figure 1. CosolvKit workflow. The input cosolvents and forcefields are passed as JSON files. When the system is created, CosolvKit outputs topology and position files that can be then simulated via MD protocols using OpenMM through CosolvKit, or using any other MD engine. Finally, CosolvKit can generate RDFs, autocorrelation plots, and density maps of the cosolvent molecules during the simulation.

During the simulation, interactions between aromatic or other highly lipophilic molecules can induce hydrophobic effects that result in aggregates in aqueous solutions[4,20]. The resulting phase separation reduces the effective concentration of the cosolvents and interferes with the sampling of the protein surface[1,21] (Figure 2a,b). When analyzing CosolvKit results, radial distribution functions of cosolvent molecules and water can be used to monitor the aggregation (Figure 2c,d). Aggregation events can be identified from anomalously high cosolvent-cosolvent density at short distances (Figure 2a, *black*) and short-range water densities that are lower than bulk water (indicating lack of a solvation shell) (Figure 2b, *black*) in plots of the radial distribution functions.

Previous methods described the application of custom repulsive forces between cosolvent molecules to avoid aggregation events[8], when running simulations with the OpenMM engine, CosolvKit allows the user to apply a repulsive force by defining the equilibrium distance and the repulsion intensity between specified system particles (as σ and ε, respectively, in Eq. 1) as the repulsive term of a Lennard-Jones potential:

$$V_{LJ}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12}\right]$$

Equation 1. Repulsive term of the Lennard-Jones potential

The application of this repulsive force between particles can restore a distribution that reflects a more uniform solvation environment (Figure 2b, and Figure 2c,d, *red*).

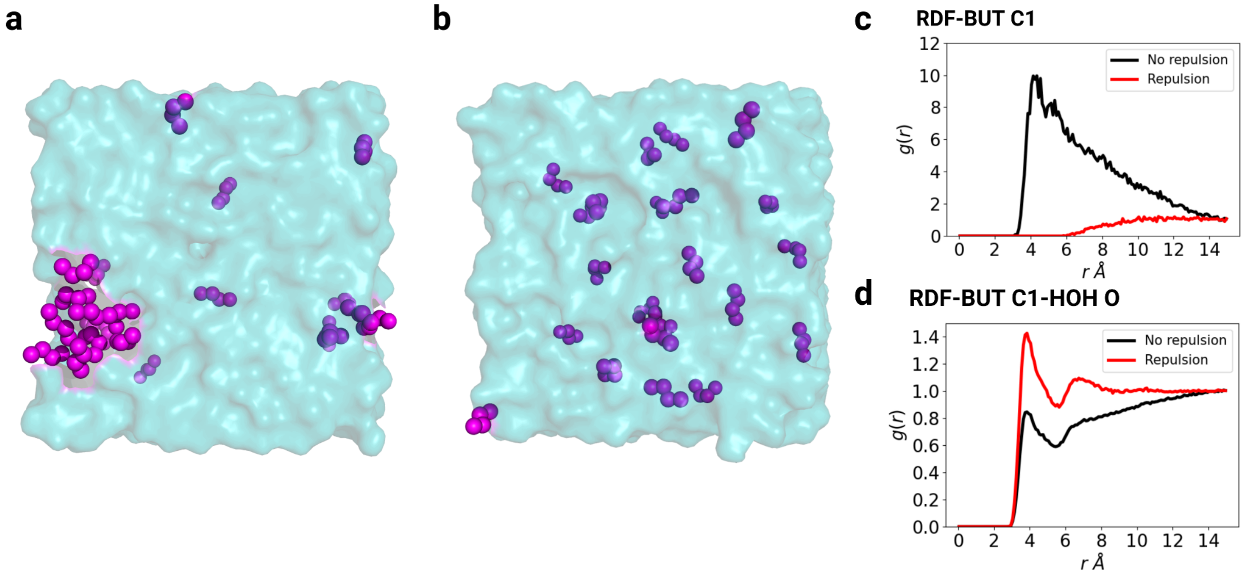

Figure 2. Cosolvent aggregation and radial distribution function of a water-butane mixture at 0.5 M (50 ns). a) Simulation of a water-butane mixture without custom repulsive forces. **b)** Simulation of a water-butane mixture with a custom repulsive term (ε=0.01 Kcal/mol, σ=5.0 Å); **c)** radial distribution function of butane-butane molecules without (*black*) and with (*red*) potential repulsion force. **d)** radial distribution function of water-butane molecules without (*black*) and with (*red*) potential repulsion force

Standard simulation protocol

A standard simulation protocol is included in the CosolvKit package, which enables the user to easily run their simulations using the OpenMM engine[11]. While certain systems may benefit from customized protocols, such as restraints on macromolecules in the initial stage of equilibration or a different water model, the defaults provided herein are expected to be reasonable for most systems.

By default, the systems are modeled with the Amber ff14sb force field[22], solvated with TIP3P-FB water molecules[23], and cosolvent probes parameters are assigned using Espaloma[12].

Langevin dynamics simulations are carried out using the Langevin Middle Integrator[24] with a temperature of 300 K and a friction coefficient of 1/ps, applying periodic boundary conditions. Electrostatic interactions are calculated using the particle mesh Ewald (PME) method[25], with a cutoff of 10.0 Å for long-range interactions, and a switching function is applied to smooth out interactions after 9.0 Å. The mass of hydrogen atoms is set to 3 u.m.a., and constraints are applied to all hydrogen bonds[26].

After the energy minimization, the system is slowly heated from 5 to 300 K in the NVT ensemble for 1000 ps with a timestep of 1 fs. Then, a Monte Carlo barostat is added to the system, and production simulations are run in the NPT ensemble, increasing the timestep to 4 fs.

A post-processing pipeline for cosolvent simulations

The main goal of CosolvKit is to provide a flexible and easy-to-use package for setting up cosolvent simulations. Developing a comprehensive solution to analyze cosolvent simulations is beyond the scope of this work and there are many tools specialized in the post-processing and analysis of cosolvent trajectories[5,8]. Nevertheless, CosolvKit provides basic post-processing and analysis functionalities to aid in the interpretation of the results.

The Report class can generate a Radial Distribution Function (RDF) for each atom in a cosolvent molecule with respect to other cosolvent copies as well as water oxygens. RDFs describe the distribution of cosolvent molecules in the simulation box and can be used to examine the solvation environment, especially to validate parameters and ensure mixing[27]. CosolvKit also plots the autocorrelation function of these RDFs. These plots reflect the time needed to sample uncorrelated distributions of solvent and cosolvent molecules, and they provide a useful metric to estimate the time required to equilibrate the cosolvent mixture. This feature is particularly helpful to identify potential issues arising from not using pre-equilibrated patches when building the systems. We used this tool to confirm that the time to achieve equilibrated cosolvent mixtures is negligible compared to standard equilibration protocols, and unlikely to impact a production run.

Taking as input a trajectory file, CosolvKit can also generate density maps for cosolvent molecules and write convenient PyMol[28] script and session files for visual inspection. Densities can be inspected individually (i.e., one for each cosolvent molecule) or together (i.e., grouped by physicochemical properties, functional groups, etc.).

To showcase the versatility and ease of use of CosolvKit, we applied it in a variety of case studies. The following section describes different applications in detail.

Case Studies

Cryptic pockets

Cryptic pockets in proteins refer to concealed binding sites that are not readily apparent in the static three-dimensional structures determined by experimental methods, but can be crucial in regulating various biological processes and are increasingly recognized as potential drug discovery targets[29–31].

The dynamic and transient nature of cryptic pockets often poses a challenge in their identification, as they are often only accessible under certain conditions or in the presence of specific ligands. In this regard, cosolvent MD simulations offer a unique advantage by facilitating the exploration of ligand-induced conformational changes[32]. However, due to slow kinetics and/or high energy barriers that separate the different conformational states, sampling limitations of conventional MD simulations can hinder its performance, requiring advanced sampling techniques to efficiently sample cryptic pocket opening events[33].

For the development of PocketMiner, Meller and colleagues compiled a dataset of proteins with experimentally confirmed cryptic pockets from the Protein Data Bank. Then, they conducted a systematic study on the ability of MD simulations to recapitulate these pockets from apo structures[2]. To demonstrate how CosolvKit can aid in the identification of cryptic pockets, we selected two test cases from their dataset and performed metadynamics cosolvent simulations to open the pockets. As a baseline, simulations were repeated for both targets using a conventional water box without any cosolvent molecule.

A commonly employed success criterion in investigations focusing on cryptic pocket discovery using MD simulations is the pocket volume[1,2], typically expressed as the fraction of the holo volume sampled during the simulation. A pocket is considered successfully opened if its volume reaches at least the volume of the holo structure. However, from a drug discovery perspective, it becomes evident that not only the volume but also the shape and distribution of polar/non-polar atoms on the pocket's surface play a pivotal role. Two pockets with identical volumes but differing shapes and/or charge distributions may not be able to accommodate the same ligand.

In addition to the pocket volume sampled during the simulations, we also calculated the accessible surface area of the pocket for every frame in the simulations and for the holo crystal using the mean *α*-sphere solvent accessibility descriptor (`mean_as_solv_acc`) in MDPocket[34].

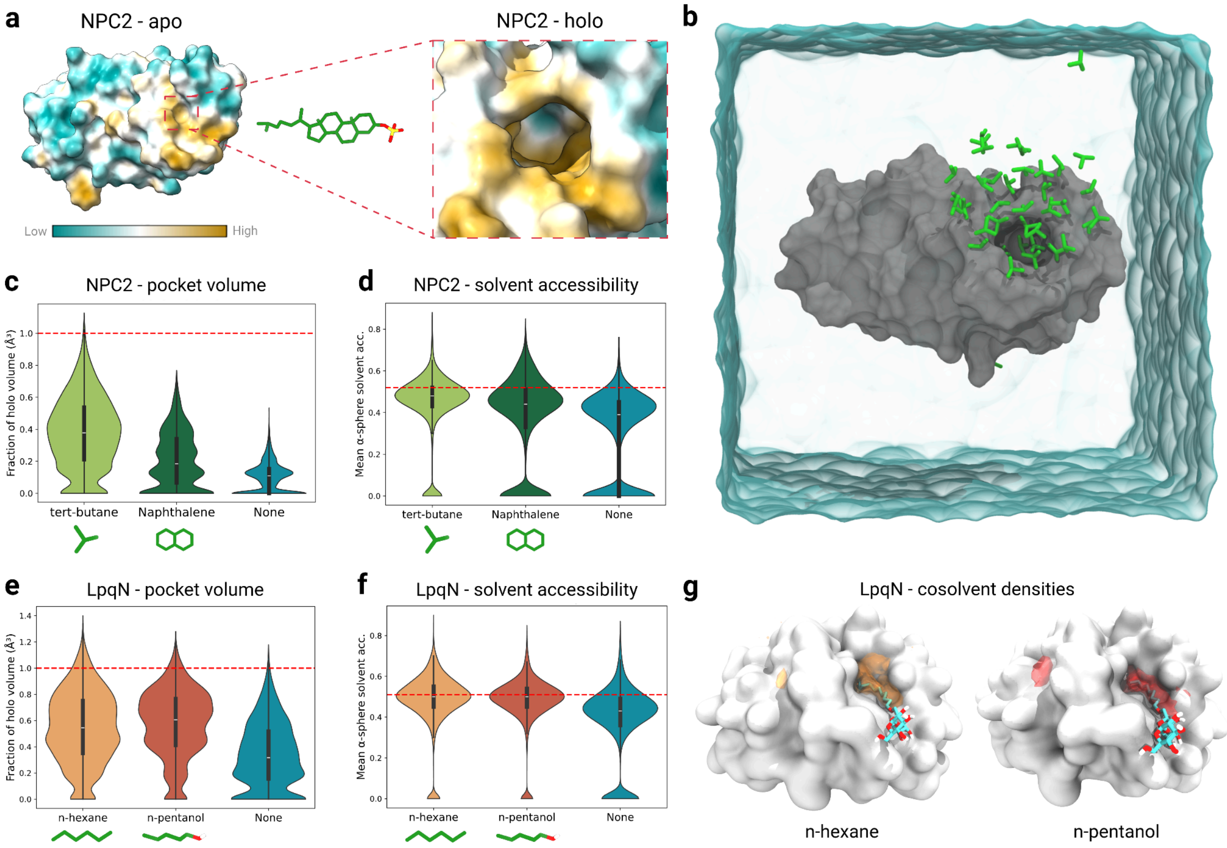

Figure 3. Opening cryptic pockets with cosolvent molecular dynamics simulations. a) surface representation of the apo (PDB id: 1nep) and holo (PDB id: 2hka) structures of NPC2, colored by hydrophobicity from low (cyan) to high (brown). The red dashed circles point to the large hydrophobic pocket that opens up upon cholesterol-3-O-sulfate binding (green sticks). **b)** a representative snapshot from the NPC2 simulation in the presence of tert-butane cosolvent probes, showing a fully open pocket with cosolvent molecules entering the cavity. **c,e)** violin plots showing the pocket volume sampled during the simulations with and without cosolvents for NPC2 and LpqN, respectively. **d,f)** violin plots for the `mean_as_solv_acc` sampled during the simulations with and without cosolvents, for NPC2 and LpqN, respectively. The red dashed lines in the violin plots represent the reference values,

calculated for the holo structure. **g)** representative structures from LpqN simulations showing the cosolvent densities calculated by CosolvKit for n-hexane (orange) and n-pentanol (red). The ligand from the holo structure was superimposed over the surface and depicted as cyan sticks.

NPC2

The bovine Niemann-Pick C2 protein (NPC2) was the only target for which the cryptic pocket could not be identified even after 5 rounds of adaptive sampling MD and 2 μs of simulation time. The authors postulated that the pocket could not be opened in water simulations due to the highly hydrophobic characteristics of the deeply buried binding site that accommodates the native ligand, cholesterol-3-O-sulfate (Figure 3a,b). Therefore, CosolvKit was used to set up two independent simulations using naphthalene and tert-butane as cosolvents.

As shown in Figure 3c, the simulations without any cosolvent were not able to sample the pocket opening according to the pocket volume. Conversely, naphthalene and *tert*-butane probes were able to shift the pocket volume distributions towards the open state, although only the former probe was able to fully sample the holo volume and open the pocket. These results further support Meller's hypothesis postulating that strong hydrophobic interactions triggered by lipophilic ligands are required to open the cryptic pocket on NPC2[2]. Notably, the *tert*-butane solvent probe not only facilitates the opening of the pocket but also samples conformations with `mean_as_solv_acc` values corresponding to solvent accessibilities more akin to the holo structure (Figure 3f).

LpqN

The second target is lipoprotein LpqN from *Mycobacterium tuberculosis*[35], which was described as a typical example of a challenging pocket opening involving secondary structure element motions and included in the validation dataset of PocketMiner. For this system, *n*-hexane and *n*-pentanol were selected as cosolvent probes.

In this case, both cosolvents, as well as the pure water simulation, were able to sample a pocket volume matching the one of the holo structure; however, cosolvent simulations achieved a higher sampling efficiency, as pocket volume was higher than in simple water simulations (Figure 3e). As shown in Figure 3g, the cosolvent density of n-hexane and n-pentanol matches the hydrophobic tale of the native ligand (PDB id: 6e5f).

Regarding the solvent accessibility surface area, while the simulation without cosolvents managed to open the pocket, the distribution of the `mean_as_solv_acc` descriptor deviates significantly from the reference value of the experimental conformation. In contrast, cosolvent probes induced an opening with properties closer to those observed in the holo pocket (Figure 3f).

Interactions mapping

To facilitate the analysis of protein binding sites that naturally interact with other proteins or peptides, CosolvKit provides a predefined minimal set of fragments to mimic the most important standard amino acid side chain features (Figure 4a). The set includes hydrophobic, hydrogen bond donors, hydrogen bond acceptors, aromatic, positively charged, and negatively charged probes (Figure 4a). This set can be extended and modified as necessary to include more specific and diverse probes as when defining other cosolvent molecules.

To demonstrate the versatility and efficacy of this peptide-focused probe set, we present 3 case studies that utilize simulations to identify determinants of protein-protein molecular recognition: Factor Xa (FxA), Major Histocompatibility Complex II (MCH II), and Lysozyme.

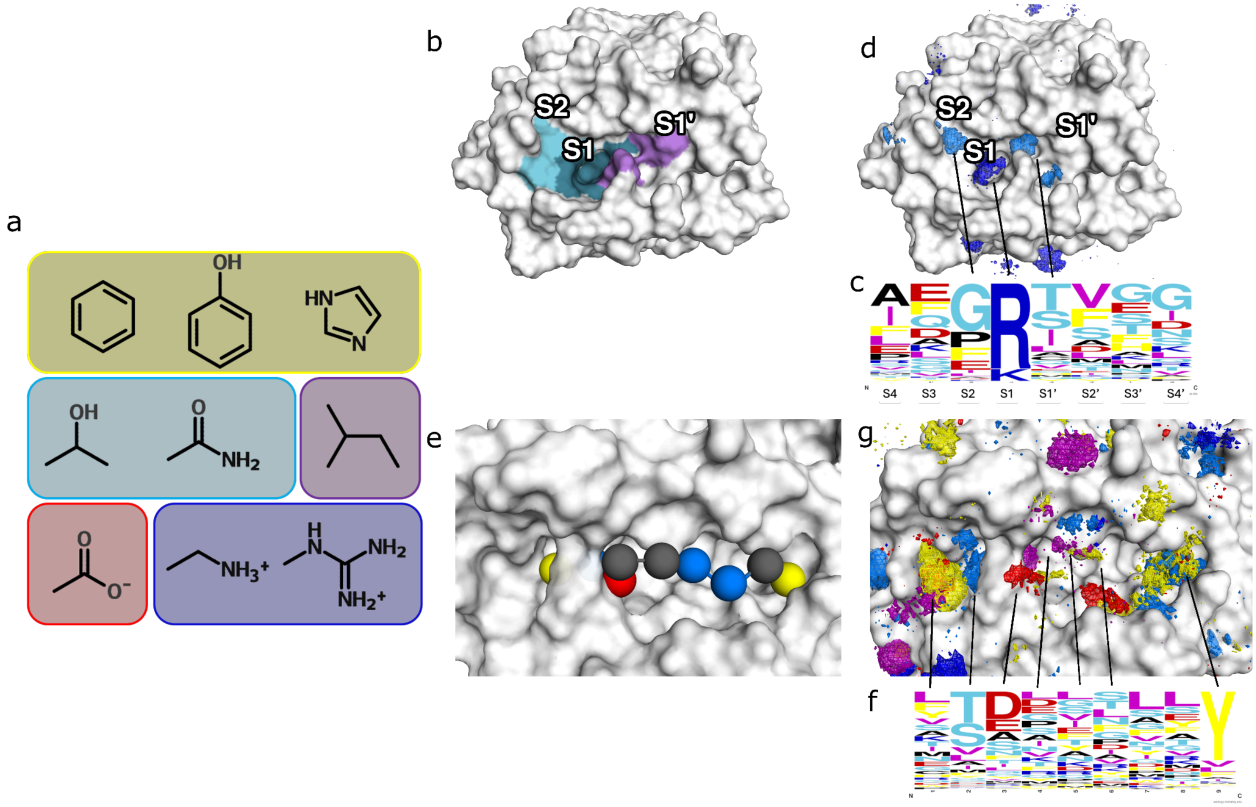

Figure 4. a) Probes used for mapping amino acid side chain interactions in the cosolvent simulation. **b)** Definition of FxA pockets according to Schechter and Berger[36] **c)** Experimental sequence logo extracted from MEROPS for FxA. Sequence logos (n=63) were logos generated using https://weblogo.berkeley.edu/[37] **d)** Grid visualization illustrating the residency probability of positively charged probes (dark blue) and hydrogen bond probes (light blue) obtained through post processing analysis of the FxA simulation (PDB id: 2P16). **e)** Property mapping of experimentally determined preferences for specific regions on MHC II. Spheres are colored by the expected binding property of respective amino acids: unspecific selection (*gray*); aromatic (*yellow*); negatively charged (*red*), hydrogen bond donor or acceptor (*light blue*); positively charged (*dark blue*); hydrophobic (*purple*). **f)** Experimentally derived sequence logo for MHC II (PDB id: 4NQX) Sequence logos (n=1062) were logos generated using https://weblogo.berkeley.edu/[37]**g)** Grid representation demonstrating the probability of residency of amino acid-mimicking probes on MHC II resulting from the CosolvKit post-processing functionality.

First, we focused on characterizing key residues essential for substrate recognition and binding of the catalytic site of FxA, a pivotal enzyme involved in the coagulation cascade[38,39]. Our primary objective was to validate interaction hotspots identified through sequence logos derived from experimental cleavage data[40]. By employing molecular dynamics simulations, we sought to elucidate the dynamic behavior of FxA and its interaction with relevant substrates or inhibitors (Figure 4b). We found a high density of positively charged probes in the S1 pocket, in agreement with the available cleavage data that reports a strong preference for arginine[40] (Figure 4c). Similarly, our analysis of the S1' pocket, known for its preference for a hydrogen bond donor moiety such as serine, revealed a significant probability of encountering one of our hydrogen bond probes, Figure 4d).

We then applied the same setup to determine critical features responsible for peptide recognition of MHC binding sites, Figure 4.e-g). Same as for proteases, experimentally-informed sequence logos are available for peptide-MHC complexes[38]. Regions demonstrating high variability in the experimental data also exhibited ambiguous densities in our simulations. The prominent aromatic profiles at the beginning and at the end of the pocket, as well as the region favoring negatively charged residues, are well represented in the simulations. In addition to mapping sites of native interactions, these probes are also appropriate, although not exhaustive, for mapping sites that can be targeted by The analysis of the interaction of these native probes can be used also to drive the identification of small molecule drugs or biotherapeutics, informing the new binding candidates identification.

For example, Factor Xa (FXa) has been extensively studied in the context of small molecule interactions[41,42], providing valuable insights into its ligand-binding properties.

The analysis of our simulations with different probes shows that they can recapitulate binding preferences and structural characteristics observed in a diverse set of crystal structures of Factor Xa[41,43,44].

Specifically, regions with high density for aromatic cosolvents, such as the S1 pocket, are in excellent agreement with the binding mode of known ligands, which tend to bind by engaging in interactions with the hydrophobic P1 pocket. Additional hydrophobic interactions can be identified with Phe99, Trp215, and Phe174. Furthermore, crystal structures have shown a consistent H-bond with Gly216[44]. Examples of such interactions are shown in Figure 5, *left*. Cosolvent densities also map to aromatic/hydrophobic features in the flatter regions of the binding site, as well as hydrogen bond features within the S2 pocket, in good agreement with crystallographic data[41].

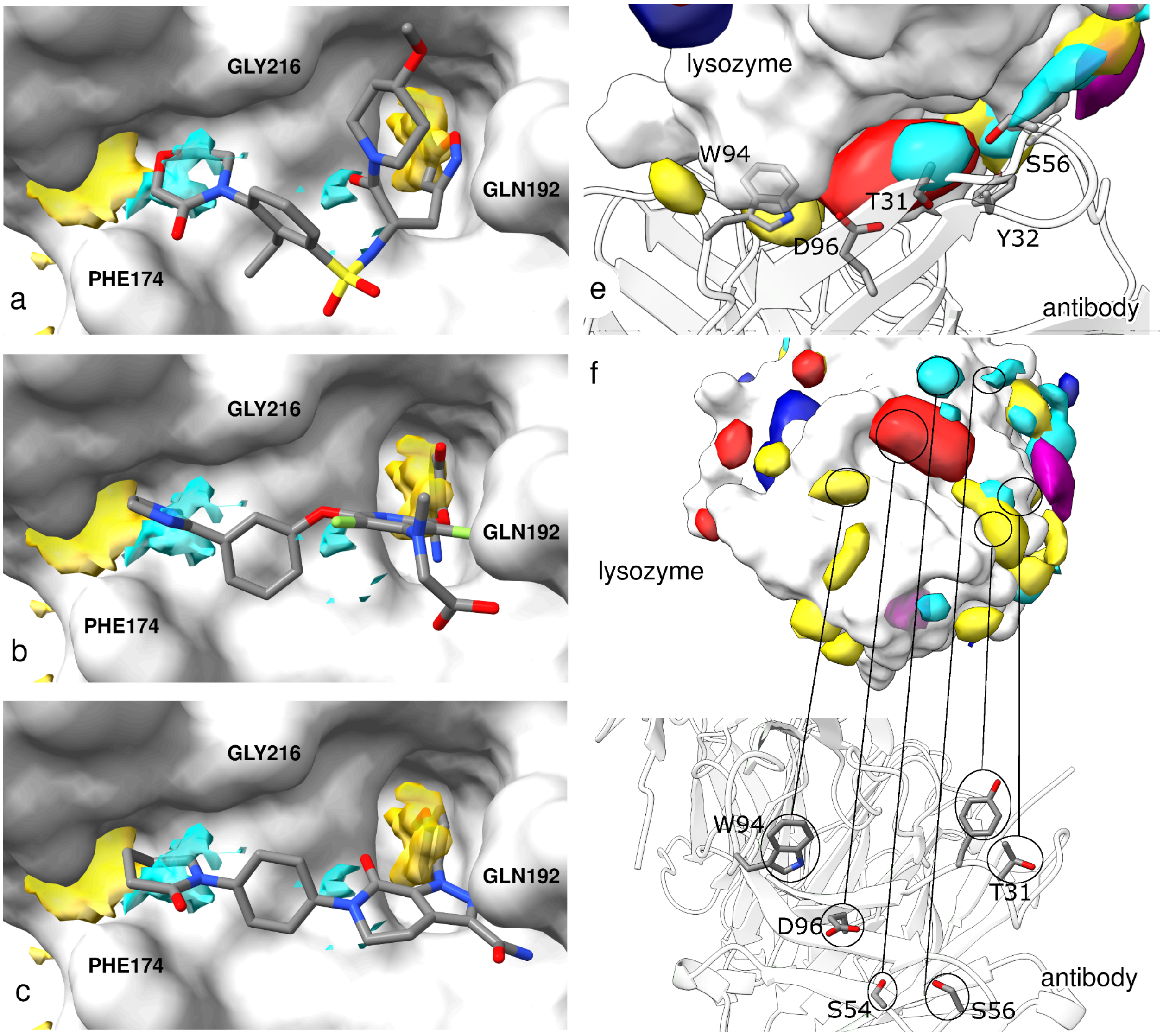

Figure 5. Overlay of simulation results showing matching pharmacophoric features with FxA crystallized ligands: **a)** PDB id: 4BTT; **b)** PDB id: 2P16; **c)** PDB id: 1FJS. **d)** Epitope interaction interface of lysozyme (surface) and antibody (cartoon and sticks) from the complex PDB id: 1P2C showing the complementarity of features and residue side chains. **e)** Cosolvent density features on lysozyme surface calculated with probes in Figure 4a (aromatic interactions yellow; hydrogen bond donor/acceptor, *light blue*; negatively charged interactions, *red;* hydrophobic interactions, *purple*) and residue side chains on the antibody surface corresponding to the key cosolvent densities identified on the Lysozyme.

Finally, we explored the possibility of using these simulations to characterize the epitope profile of antibodies using lysozyme as prototypical antigen[41].

Cosolvent simulations were performed starting from a lysozyme structure originally crystallized with the respective antibody[45]. Previous studies have shown that the epitope shows reduced flexibility for this antigen, facilitating the characterization of the epitope properties[46]. Simulation analysis showed a good agreement between the residues involved in antigen binding and the corresponding patches identified through cosolvent analysis (Figure 5). Regions of the epitope favoring an aromatic probe accurately matched with tryptophan or tyrosine residues in the crystal structure while regions with a preference for the negative probe matched with the presence of aspartate residues in the paratope stabilizing the interactions with the lysozyme antigen.

Conclusion

CosolvKit was designed to provide a flexible platform to configure, run, and analyze complex cosolvent simulations. This toolkit expands on limitations in existing methods by allowing user-defined cosolvents and concentrations. This flexibility is essential for cosolvent simulations, where the concentration and identity of cosolvents may have large impacts on the results of simulations. Importantly, CosolvKit is flexible in its use, exposing functionality either from the command line interface or the Python API, simplifying its integration with the OpenMM environment. Finally, reproducibility is encouraged by the use of a configuration file that captures the essential parameters necessary to run consistent simulations across targets and experiments.

Through a series of case studies, we demonstrate the broad applicability of CosolvKit. These simulations map the sites of native and designed ligands for therapeutically relevant proteins, and the opening of otherwise intractable cryptic binding pockets. These simulations also demonstrate the versatility of our pipeline, where systems prepared with CosolvKit may be subjected to diverse molecular dynamics simulation protocols, including the use of enhanced sampling techniques. We showed how this method can be effectively used to drive structure-based drug design, including antibody-antigen engineering.

Notably, most of these simulations would be particularly challenging to set up with existing cosolvent pipelines due to their limitations in the nature and number of probes, their concentrations, or both. Furthermore, while CosolvKit contains basic analysis and reporting tools to monitor cosolvent solvation and report densities, it is designed to be extensible and easily integrated into other analysis pipelines. We hope the ease of use and flexibility of CosolvKit will provide a platform for further scientific development, including the investigation of diverse cosolvent mixtures and novel methods.

Acknowledgments

This work has been supported by NIH grants R24-GM145962 and R01-GM069832 from NIGMS (SF).

Contributions

N.B. designed the project and developed the software. J.E. conceived the project and contributed to the software development. M.L. contributed to the software development and performed and analyzed MD simulations. J.R.L. and M.L.F.Q. performed and analyzed MD simulations. M.H. and D.S.M. provided critical input and contributed to the software development. All authors contributed to the preparation of the manuscript. A.B.W. and S.F. provided funding and resources to support the project.

Code availability

CosolvKit is released as open source under a LGPL2.1 license. The source code, the documentation, and updates are available on GitHub: https://github.com/forlilab/cosolvkit/releases.

1. Oleinikovas, V., Saladino, G., Cossins, B. P. & Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J. Am. Chem. Soc.* **138**, 14257–14263 (2016).
2. Meller, A. *et al.* Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* **14**, 1177 (2023).
3. Schmidt, D., Boehm, M., McClendon, C. L., Torella, R. & Gohlke, H. Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. *J. Chem. Theory Comput.* **15**, 3331–3343 (2019).
4. Szabó, P. B., Sabanés Zariquiey, F. & Nogueira, J. J. Cosolvent and Dynamic Effects in Binding Pocket Search by Docking Simulations. *J. Chem. Inf. Model.* **61**, 5508–5523 (2021).
5. Sabanés Zariquiey, F., de Souza, J. V. & Bronowska, A. K. Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome. *Sci. Rep.* **9**, 19118 (2019).
6. Graham, S. E., Leja, N. & Carlson, H. A. MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. *J. Chem. Inf. Model.* **58**, 1426–1433 (2018).
7. Alvarez-Garcia, D. & Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **57**, 8530–8539 (2014).
8. Faller, C. E., Raman, E. P., MacKerell, A. D. & Guvench, O. Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design. in *Fragment-Based Methods in Drug Discovery* (ed. Klon, A. E.) 75–87 (Springer, New York, NY, 2015). doi:10.1007/978-1-4939-2486-8_7.
9. Yanagisawa, K., Moriwaki, Y., Terada, T. & Shimizu, K. EXPRORER: Rational Cosolvent Set Construction Method for Cosolvent Molecular Dynamics Using Large-Scale Computation. *J. Chem. Inf. Model.* **61**, 2744–2753 (2021).
10. Kimura, S. R., Hu, H. P., Ruvinsky, A. M., Sherman, W. & Favia, A. D. Deciphering Cryptic

Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model.* **57**, 1388–1401 (2017).

11. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).
12. Wang, Y. *et al.* End-to-end differentiable construction of molecular mechanics force fields. *Chem. Sci.* **13**, 12016–12033 (2022).
13. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
14. Boothroyd, S. *et al.* Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **19**, 3251–3275 (2023).
15. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
16. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
17. Brooks, B. R. *et al.* CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
18. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370-376 (2012).
19. Kocis, L. & Whiten, W. J. Computational investigations of low-discrepancy sequences. *ACM Trans. Math. Softw.* **23**, 266–294 (1997).
20. Thomas, M. *et al.* Atomistic simulations of the aggregation of small aromatic molecules in homogenous and heterogenous mixtures. *Phys. Chem. Chem. Phys. PCCP* **22**, 21005–21014 (2020).
21. Guvench, O. & Jr, A. D. M. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLOS Comput. Biol.* **5**, e1000435 (2009).

22. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

23. Wang, L.-P., Martinez, T. J. & Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett.* **5**, 1885–1891 (2014).

24. Zhang, Z., Liu, X., Yan, K., Tuckerman, M. E. & Liu, J. Unified Efficient Thermostat Scheme for the Canonical Ensemble with Holonomic or Isokinetic Constraints via Molecular Dynamics. *J. Phys. Chem. A* **123**, 6056–6079 (2019).

25. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

26. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–508 (2001).

27. Lexa, K. W., Goh, G. B. & Carlson, H. A. Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations. *J. Chem. Inf. Model.* **54**, 2190–2199 (2014).

28. The PyMOL Molecular Graphics System, Version 2.5 Schrödinger, LLC.

29. Comitani, F. & Gervasio, F. L. Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. *J. Chem. Theory Comput.* **14**, 3321–3331 (2018).

30. Zimmerman, M. I. *et al.* SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* **13**, 651–659 (2021).

31. Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M. & Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **44**, 1–8 (2018).

32. Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc. Chem. Res.* **53**, 654–661 (2020).

33. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly

Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100**, 020603 (2008).

34. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
35. Melly, G. C. *et al.* Structural and functional evidence that lipoprotein LpqN supports cell envelope biogenesis in Mycobacterium tuberculosis. *J. Biol. Chem.* **294**, 15711–15723 (2019).
36. Schechter, I. & Berger, A. On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).
37. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
38. Camire, R. M. Blood Coagulation Factor X: Molecular Biology, Inherited Disease, and Engineered Therapeutics. *J. Thromb. Thrombolysis* **52**, 383–390 (2021).
39. Brown, M. A., Stenberg, L. M. & Stenflo, J. Chapter 642 - Coagulation Factor Xa. in *Handbook of Proteolytic Enzymes (Third Edition)* (eds. Rawlings, N. D. & Salvesen, G.) 2908–2915 (Academic Press, 2013). doi:10.1016/B978-0-12-382219-2.00642-6.
40. Rawlings, N. D., Waller, M., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **42**, D503–D509 (2014).
41. Meneyrol, J. *et al.* 5-Chlorothiophene-2-carboxylic Acid [(S)-2-[2-Methyl-3-(2-oxopyrrolidin-1-yl)benzenesulfonylamino]-3-(4-methylpiperazin-1-yl)-3-oxopropyl]amide (SAR107375), a Selective and Potent Orally Active Dual Thrombin and Factor Xa Inhibitor. *J. Med. Chem.* **56**, 9441–9456 (2013).
42. Komoriya, S. *et al.* Design, synthesis, and biological activity of non-basic compounds as factor Xa inhibitors: SAR study of S1 and aryl binding sites. *Bioorg. Med. Chem.* **13**, 3927–3954 (2005).
43. Adler, M. *et al.* Preparation, Characterization, and the Crystal Structure of the Inhibitor

ZK-807834 (CI-1031) Complexed with Factor Xa,. *Biochemistry* **39**, 12534–12542 (2000).

44. Pinto, D. J. P. *et al.* Discovery of 1-(4-Methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro-1H-pyrazolo[3,4-c]pyridine-3-carboxamide (Apixaban, BMS-562247), a Highly Potent, Selective, Efficacious, and Orally Bioavailable Inhibitor of Blood Coagulation Factor Xa. *J. Med. Chem.* **50**, 5339–5356 (2007).
45. Cauerhff, A., Goldbaum, F. A. & Braden, B. C. Structural mechanism for affinity maturation of an anti-lysozyme antibody. *Proc. Natl. Acad. Sci.* **101**, 3539–3544 (2004).
46. Fernández-Quintero, M. L. *et al.* Conformational selection of allergen-antibody complexes-surface plasticity of paratopes and epitopes. *Protein Eng. Des. Sel. PEDS* **32**, 513–523 (2019).

Methods

CosolvKit - post processing densities generation

The densities of cosolvent molecules are calculated with the use of the python packages numpy and griddataformats. First, the multidimensional histogram of the cosolvent positions during the simulation is computed with the histogramdd module which is then used to instantiate a Grid class instance to create the density map of each cosolvent molecule for the whole simulation.

Custom forces simulations

The simulations to test the custom repulsive force application were carried out with the standard procedure, using the prepare_cosolvent_system.py script. The cosolvent used was butane at 0.5 M concentration parametrized with the espaloma force field and solvated in a cubic water box of 12 Å. The simulations were set up with amber14 and tip3pfb force fields and were run for 50ns with the OpenMM MD engine.

When applied, the custom repulsive forces were specified between butane atoms of different molecules with ε=0.01 Kcal/mol and σ=5.0 Å.

Cryptic pockets

Apo structures of NPC2 (PDB id: 1nep) and LpqN (PDB id: 6e5f) were obtained from the Protein Data Bank and prepared using PDBFixer, repairing missing segments and heavy atoms, keeping all crystallographic waters, and adding hydrogens according to a physiological pH of 7.4. Considering a padding distance of 12 Å around each structure, a cubic box was constructed and filled up with tip3pfb water molecules and the selected cosolvent, to reach a concentration of 0.5 M. Sufficient NaCl was added to neutralize the systems.

For each system, 10 independent trajectories or walkers were simulated for 200 ns each, using OpenMM implementation of well-tempered metadynamics algorithm(see ref 33 in the main text) . The root-mean-square deviation (RMSD), calculated over all protein heavy atoms, was chosen as the collective variable to boost during

https://doi.org/10.26434/chemrxiv-2024-rmsnj **ORCID:** https://orcid.org/0000-0003-0223-3214

metadynamics simulations. The bias was deposited every 500 ps, with a sigma of 0.5 Å and a hill height of 0.3 Kcal/mol. The bias factor was set to 15 for all the runs. All other simulation parameters were set as described in the simulation protocol section.

FxA, MHCII and Lysozyme

Structures of FxA (PDB id: 2P16) and MHCII (PDB id: 4NQX) and lysozyme (PDB id: 1p2c) were obtained from the Protein Data Bank. For FxA and MHCII bound ligands were removed. For lysozyme, the antibody was removed to expose the antigen. PDBFixer was used to prepare the structures, keeping all crystallographic waters, and adding hydrogens according to a physiological pH of 7.4. Considering a padding distance of 10 Å around each structure, a cubic box was constructed and filled up with tip3pfb water molecules and the mix of amino acid mimicking cosolvent, set to a concentration of 0.1 M respectively. Sufficient NaCl was added to neutralize the systems.

For each system, 5 independent trajectories or replicas were simulated for 100 ns each, using OpenMM. All other simulation parameters were set as described in the simulation protocol section.