Functionally Important Residues from Graph Analysis of Coevolved Dynamic couplings

Manming Xu,[1*] Sarath Chandra Dantu,[2*] Robert A Bonomo,[3,4,5,6,7,8] James A Garnett,[9] Alessandro Pandini ,[2,10^] and Shozeb Haider[1,10^]

[1]UCL School of Pharmacy, London WC1N 1AX UK
[2]Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, UK
[3]Research Service, Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, OH, USA;
[4]Department of Molecular Biology and Microbiology, Case Western Reserve University School of Medicine, Cleveland, OH, USA;
[5]Department of Medicine, Case Western Reserve University School of Medicine, Cleveland, OH, USA;
[6]Clinician Scientist Investigator, Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, OH, USA;
[7]Departments of Pharmacology, Biochemistry, and Proteomics and Bioinformatics, Case Western Reserve University School of Medicine, Cleveland, OH, USA;
[8]CWRU-Cleveland VAMC Center for Antimicrobial Resistance and Epidemiology (Case VA CARES) Cleveland, OH, USA;
[9]Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London, UK
[10]The Thomas Young Centre for Theory and Simulation of Materials, London SW7 2AZ, UK
* Joint first authors

^ To whom correspondence should be addressed:
Prof Shozeb Haider (Shozeb.haider@ucl.ac.uk)
Dr Alessandro Pandini (alessandro.pandini@brunel.ac.uk)

ORCID:
Shozeb Haider: 0000-0003-2650-2925
Alessandro Pandini: 0000-0002-4158-233X

Abstract

Proteins perform substrate binding and catalysis, aided by conformational dynamics. A mixture of short- and long-range inter residue interactions are responsible for orchestrating both structural changes and the biological activity. Hence, it is critical to understand the relationship between protein dynamics and function and identify the responsible residues, as this information can be crucial for developing effective therapeutics or for altering protein functions for biotechnological purposes. Existing computational methods for the predictions of functional residues are trained on sequence, structural and experimental data, but they do not explicitly model the influence of evolution on protein dynamics. This overlooked contribution needs to be considered as it is known that evolution can fine tune protein dynamics through compensatory mutations, either to improve the proteins performance or diversify its function maintaining the same fold scaffold. To model this critical contribution, we introduce DyNoPy, a computational tool that combines residue coevolution analysis with molecular dynamics (MD) simulations and using network analysis it pinpoints important residues for functional dynamics. We applied DyNoPy to an example of class-A and class-C β-lactamases. SHV-1 with its rich experimental data serves as a test case with known functional annotations, while PDC-3 is used to show the method's predictive power on a less characterised protein. DyNoPy successfully identifies residues with annotated function and facilitates the explanation of mutations at previously unexplored sites. The results have significant potential for investigation of new clinically relevant β-lactamase variants and confirms the potential of DyNoPy to inform drug design strategies.

Significance

Protein function is mediated by structural dynamics, and it is now accepted that mutation events can directly impact dynamical properties. To model and predict these effects, it is essential to examine interactions at the residue level. Molecular simulations and sequence coevolution analysis separately offer insights into physical interactions and evolutionary changes. DyNoPy is an integrated computational method to discover coevolved residue pairs that contribute to protein functional dynamics. These residues are represented as a network where specific groups of nodes are coupled with protein function. Unveiling the critical roles of these residues and their dynamic interactions is shown to explain pathogenic effects of mutations and can be used to guide rational design of protein function.

Introduction

Quantifying the contribution of individual residues or residue groups to protein function is important to estimate the pathogenic effect of mutations (1). Identifying the functional roles of individual residues has primarily been done through mutagenesis experiments (2). Bioinformatics methods have complemented these approaches through analysis of multiple sequence alignments (MSA) of homologous proteins and structural data (3-8). Among these methods, computational techniques that can decode inter-residue evolutionary relationships from MSAs have paved the way for machine learning (ML) based strategies that can predict protein structure (9-12), stability (13), and function (7) and extend the scope of computational protein design (14-16). A most recent approach has combined experimental data from three proteins on stability and function with sequence and structural features to train a ML model to predict functional sites (17).

Functional sites are often regulated by both local and global interactions. Changes in these interactions are instrumental for functional events like substrate binding, catalysis, and conformational changes (18). While techniques described earlier have exploited the sequence and structure relationship of proteins, they do not directly consider the time evolution of functional events in their predictions. The development of physical models of protein dynamics and the increase in available computational power has stimulated the adoption of computational techniques (19, 20) to investigate the conformational dynamics of proteins, an essential component of the many biological functions (21, 22). At the same time, it has been established that evolution through compensatory mutations in dynamics regions, like hinges and loops, can fine tune protein dynamics for function and introduce promiscuity in dynamics, diversifying biological function. Assuming that protein functional dynamics is conserved during evolution, significant information on dynamic regions and substrate recognition sites should be recoverable using inter residue coevolution scores extracted from MSAs (23, 24). Coevolution analysis and Molecular Dynamics (MD) simulations have independently (25) and synergistically been combined in the past to identify important residues for function (26-30). Yet a tool that combines hidden information on dynamics from evolution with direct information on local and global dynamics from conformational ensembles from MD is not available.

We present DyNoPy, a computational method that can extract hidden information on functional sites from the combination of pairwise residue coevolution data and powerful descriptors of dynamics extracted from the analysis of MD ensembles. The method can detect coevolved dynamic couplings, i.e. residue pairs with critical dynamical interactions that have been preserved during evolution. These pairs are extracted from a graph model of residue-residue interactions. Communities of important residue groups are detected, and critical sites are identified by their eigenvector centrality in the graph. A comprehensive overview of the method is shown in Figure 1. Two β-lactamases from distinct families, SHV-1 from class A and PDC-3 from class C (31, 32), were used to demonstrate potential use of DyNoPy. In these cases, DyNoPy can successfully detect residue couplings that align with previous studies, guide in the explanations of mutation sites with previously unexplained mechanisms and provide predictions on plausible important sites for the emergence of clinically relevant variants.

Results and Discussion

β-lactamases are a group of enzymes capable of hydrolysing β-lactams, conferring resistance to β-lactam antibiotics (33). These enzymes are evolving rapidly, as single amino acid substitutions are sufficient to drive their evolution and increase their catalytic spectrum and inhibitor resistance profile (34). Extended-spectrum β-lactamases (ESBLs) and inhibitor resistant β-lactamases (IRBLs) are frequently identified from clinics, highlighting their global impact on antibiotic resistance (35). The rapid evolution of β-lactamase proteins and their clinical significance (34) makes them an ideal target for assessing the robustness of DyNoPy.

We have applied DyNoPy to two model enzymes from different β-lactamase families: class A β-lactamase SHV-1 and class C β-lactamase PDC-3 (31, 32) (Figures S1 and S2). Both class A and class C β-lactamases comprise an α/β domain and an α helical domain, with the active site situated in between (36, 37). Moreover, both enzymes target the carbonyl carbon of the β-lactams using a serine residue (38, 39). Despite these similarities, the structures of class A and class C β-lactamases are remarkably different (Figure 2).

SHV-1 is a very well characterised protein system with wealth of information on mutations and their corresponding effects on protein function and in contrast PDC-3 remains to be characterised. Detailed structural information on these proteins can be found in the supplementary materials. Essential catalytic residues in SHV-1 are: S_{70}, K_{73}, S_{130}, E_{166}, N_{170}, K_{234}, G_{236}, and A_{237} (40) and conserved catalytic residues in PDC-3 include S_{64}, K_{67}, Y_{150}, N_{152}, K_{315}, T_{316}, and G_{317}. Highly conserved stretches of 3-9 hydrophobic residues, annotated as hydrophobic nodes, exists in class A β-lactamases and have been proven to be essential for protein stability (41). Residues defined as belonging to hydrophobic nodes within SHV-1 are listed in Supplementary Table S1.

In SHV-1, the predominant ESBLs substitutions occur at L_{35}, G_{238}, and E_{240}, R_{43}, E_{64}, D_{104}, A_{146}, G_{156}, D_{179}, R_{202}, and R_{205} appear in ESBLs with lower frequency (42). Mutations at M_{69}, S_{130}, A_{187}, T_{235}, and R_{244} are known to induce inhibitor resistance in the enzyme (43). In PDC-3, substitutions primarily occur on the Ω-loop, enhancing its flexibility to accommodate the bulky side chains of antibiotics, while deletions are more common in the R2-loop (39). The predominant Ω-loop mutations isolated from clinics are found at positions V_{211}, G_{214}, E_{219}, and Y_{221} (44).

Emergence of highly conserved dynamic couplings

DyNoPy builds a pairwise model of conserved dynamic couplings detected by combining coevolution scores and information on functional motions into a score J_{ij} (see Methods and Fig. 1). To this end a dynamics descriptor should be selected. When the descriptor is associated with functional conformational changes, it is expected that functionally relevant couplings will report higher scores. Dynamics descriptors can be selected from commonly used geometrical collective variables (CVs) for the analysis of MD trajectories (see Methods). As expected, the average $\boldsymbol{J}$ matrix score varies across the different CVs, with some of them showing no signal of dynamic coupling (Supplementary Fig. S4C).

SHV-1 and PDC-3 exhibit distinct dynamics, requiring a different choice of the CV that best captures the functional dynamics. For SHV-1, the global first principal component (PC1) proved to be the most effective feature, identifying 571 residue pairs with a J_{ij} value greater than 0. Conversely, PDC-3 requires selection of more localized features that can extract the Ω-loop dynamics from the overall protein motion. Among the dynamic descriptors, the partial first time-lagged component (TC1_partial) performed best for PDC-3, detecting 216 residue pairs with a J_{ij} value greater than 0. Consequently, PC1 and TC1_partial were selected to build the $\boldsymbol{J}$ matrix for SHV-1 and PDC-3, respectively. The performance of all 12 CVs for each protein was assessed and listed in the Supplementary Table S2.

The importance of dynamical information is evident when coevolution couplings (γ_{ij}) and conserved dynamic couplings (J_{ij}) are compared: the number of non-zero couplings decrease from 40% to <2% of total residue pairs in the protein (Fig. S4D) when information from the dynamics descriptor is added. Thus, the inclusion of protein dynamics in coevolution studies acts as an effective filter that rules out residue pairs that do not have significant correlations with functional motions. Moreover, when relying only on γ_{ij}, all the residues in SHV-1 and PDC-3 are included within four identified communities (Supplementary Table S3), suggesting that coevolution scores (γ_{ij}) alone do not effectively discriminate residues relevant for protein functions. Furthermore, it would be hard to distinguish critical core residues for each community using only γ_{ij}, as the eigenvector centrality (EVC) values for the residues do not show remarkable differences (Supplementary Figures S5A and S5B). This means that a detailed investigations of the top

residues is needed to determine which pairs should be picked up and further analysed. On the other hand, it is much easier to identify essential residues based on J scores calculated, as clear outliers with significantly higher EVC values could be seen for almost all communities (Supplementary Figures S5C and S5D) (25, 45). In conclusion, the lack of specificity in the statistically based coevolution analysis supports the choice of incorporating a score for the correlation between residue interactions and dynamic behaviours.

DyNoPy reveals critical residues and predicts evolutionary pathways in SHV-1

DyNoPy identified eight significant communities of strongly coupled residues within SHV-1 (Supplementary Figure S4A) with all crucial catalytic residues and critical substitution sites previously mentioned participating in one of these communities with the exceptions of R_{43}, R_{202}, and S_{130}. Residues previously known to have critical role in function or conferring ESBLs/IRBLs phenotype are either directly coupled to protein dynamics or act as a central hub. The hubs interact with residues with either a role in catalysis or structural stability through their membership of hydrophobic nodes (31). Furthermore, DyNoPy identified key positions (L_{162} and N_{136}) within some communities that are known to undergo substitutions, conferring an ESBL phenotype in other class A β-lactamases. These substitutions have not yet emerged in the SHV family, providing insightful predictions about the potential future evolution of the enzyme. Detailed description of the other three communities is provided in the supplementary information (Figure S6).

DyNoPy predicts mutation hotspots for SHV-1

DyNoPy detects critical mutation sites (L_{162} and N_{136}) that are known to extend the range of possible substrates in other class A β-lactamases, but have not yet emerged as variants in the SHV family . These sites have not been modified in SHV family because of their plausible central role within the communities as they are mediating couplings with key functional residues essential for catalytic activity and structural stability, indicating their critical role in protein function and the potential lower mutation rate. These findings provide insightful predictions about the potential future evolution of the enzyme, as well as plausible explanations for why these mutations have not yet appeared.

$L_{162,}$ positioned at the start of the Ω-loop and adjacent to the crucial catalytic residue $E_{166,}$ is assigned as the core residue for community 1 (Figure 3A). While it remains conserved in SHV family, variants of L_{162} are known in other class A β-lactamase that are known to expand the enzyme catalytic spectrum. Single amino acid substitution at L_{162} can intensify antibiotic resistance in BEL-1 (46), a class A ESBL clinical variant, exhibiting robust resistance to ticarcillin and ceftazidime (47). BEL-2 diverges from BEL-1 by single amino acid substitution ($L_{162}F$) which alters the kinetic properties of the enzyme significantly and increases its affinity towards expanded-spectrum cephalosporins (48). The relationship between L_{162} and protein catalytic functions can be explained using DyNoPy model, as there are couplings with catalytic important residues M_{69}, K_{73}, E_{166}, and K_{234}. Moreover, the BEL case has confirmed that $L_{162}F$ mutation significantly destabilizes the overall protein structure, highlighting the crucial role of L_{162} in maintaining protein stability (46). DyNoPy accurately identifies the centrality of L_{162} by reporting its connections with 28 backbone residues, including nine hydrophobic node residues critical for protein stability. Among these, five hydrophobic residues are part of the α2 node: V_{75}, L_{76}, G_{78}, V_{80}, and L_{81}, highlighting the contribution of L_{162} to the stability of the α2 helix (31).

Just like L_{162}, N_{136} undergoes advantageous mutations in other class A β-lactamases while remains highly conserved within the SHV family. It is the core residue for community 7 (Figure 4B). This residue forms a hydrogen bond with E_{166}, stabilizing the Ω-loop (49). N_{170} acts as an intermediary between N_{136} and E_{166}. N_{170}, an essential catalytic residue located on the Ω-loop, contributes to priming the water molecule for the deacylation of β-lactams with E_{166} (50) and is directly coupled with N_{136}. Due to the essential contribution of N_{136} in facilitating E_{166} to maintain its proper orientation, it was previously thought to be intolerant to mutations as substitution of Asparagine to Alanine at this position would make the enzyme lose its function completely (51). However, $N_{136}D$ substitution has emerged as a new clinical variant very recently in PenL, a class A β-lactamase, by increasing its ability in hydrolysing ceftazidime (51), suggesting that this site has potential to mutate. This gain of function is mainly triggered by the increased flexibility of the Ω-loop (51). DyNoPy correctly detect a dynamical relationship between N_{136} and the Ω-loop (residues 164-179). Six residues present in the Ω-loop participate within this community, including R_{164} and D_{179}. These two residues are critical as they are forming the 'bottleneck' of the

Ω-loop which is essential for the correct position of E_{166} (52). D_{179} is also a critical mutation site for SHV-1. Single amino acid substitutions like $D_{179}A$, $D_{179}N$, and $D_{179}G$ are enough for the extended spectrum phenotype (42).

DyNoPy detects residue couplings essential for protein stability

DyNoPy identifies residue couplings critical for protein functional motions, particularly associated with protein stability. These residue pairs exhibit strong relationships as they are not only directly coupled with each other, but also forms various indirect couplings via other residues. As a result, both residues are considered as core residues inside these communities. It is expected that disruption of these couplings through mutation could compromise collective motions essential for enzyme activity.

As the secondary core residues in community 1 (Figure 3A) F_{72} is showing a strong coupling with the primary core residue L_{162} and also forms nine indirect couplings with L_{162}, including via the catalytic K_{234}. This network of direct and indirect relationships reveals the importance of F_{72} and L_{162} coupling in maintaining protein functional motions. Interestingly, previous studies identified a small hydrophobic cavity formed by L_{162} and F_{72}, together with L_{139}, and L_{148}, which is essential for the stability of the active site (46). Notably, DyNoPy successfully recovers the key residues of this local hydrophobic cavity (L_{162}, F_{72}, and L_{148}).

The strong interplay between V_{103} and S_{106}, which are both residues on the α3-α4 loop, is seen in community 5 (Figure 3C). These residues not only interact with each other directly but are also indirectly coupled via 21 other residues. This community emphasizes the significance of hydrophobic nodes in SHV stability and dynamics. Within the analysed 48 residues, 27 are hydrophobic, out of which 15 residues act as nodes critical for enzyme stabilization. Hydrophobic nodes stabilize their own secondary structures and interconnect to stabilize the overall protein (53). V_{103} and S_{106} themselves are hydrophobic nodes, stabilizing α3 helix and α4 helix respectively, and are strongly coupled with each other. In CTX-M, another class A enzyme, $N_{106}S$ is a common substitution that results in improved thermodynamic stability and compensate for the loss in stability of the variants (54). Interestingly, this residue is already a Serine in SHV, but still implies its pivotal role in protein stability.

DyNoPy provides valid explanations for mutation sites

During the evolution of β-lactamases, single mutations on specific sites that are distant from the functional sites have been observed to significantly alter protein catalytic functions. Additionally, single mutations on some surface exposed residues can dramatically increase protein stability. Understanding how these distant mutations impact function and stability becomes a major challenge in understanding protein evolutionary pathways. Communities extracted by DyNoPy show these residues linked with functional important residues, providing a rational for these mutation sites with unknown functions.

Mutations of G_{156} are limited but they lead to ESBL phenotype in the SHV family (42). G_{156} is the central residue for community 4 (Figure 3B), but it is distant from the active site, over 20 Å away from the catalytic serine S_{70}. Clinical variant SHV-27, has extended resistance ability towards cefotaxime, ceftazidime, and aztreonam (55). It differs from SHV-1 by single amino acid substitution $G_{156}D$, suggesting that it has directly evolved from SHV-1 (55). Limited research has been done on position G_{156}, and the understanding of how it affects the enzyme catalytic properties given that it is far away from the active site is still unclear. Based on our results, we suggest that this residue is essential for the overall protein function because of its 11 coevolved dynamic couplings with protein dynamics, including A_{146}, another ESBL substitution site.

SHV-38, an ESBL that is capable of hydrolysing carbapenems, harbours a single $A_{146}V$ substitution compared to SHV-1 (56). Like G_{156}, A_{146} is 15 Å away from S_{70} but shows an ability in altering protein catalytic function. The A_{146}-G_{156} residue pair shows a strong coevolutionary signal and strong correlation with protein overall dynamics, implying that there may compensatory mutations at these sites with potential to emerge in the SHV family in the future. These two residues are not connected to any catalytic residues but their coupling to functional dynamics can offer plausible explanation to ESBL activity of these two mutations.

Unlike other substitution sites that are adjacent to the active site, R_{205} is situated more than 16 Å away from catalytic serine S_{70}. Its side chain points outward from the protein, exposing to the solvent. The $R_{205}L$ substitution often co-occurs with other ESBL mutations and is thought to indirectly contribute to the ESBL phenotype by compensating for stability loss induced by other mutations (57). SHV-3 is an ESBL that exhibits significant resistance to cefotaxime and

ceftriaxone (58). Two substitutions in this enzyme, $R_{205}L$ and $G_{238}S$, extend its resistance profile (58). Thus, it is promising to see that DyNoPy detected these two mutation sites together within community 6 (Figure 4A).

Y_{105} and R_{266} are the core residues for community 6. Y_{105} is situated on the α3-α4 loop positioned at the left side of the binding pocket. It is an important catalytic residue that recognizes and binds to the thiazolidine ring of penicillins or β-lactamase inhibitors (59). There is very limited information on the role of R_{266}, except that it may stabilize the Ω-loop in the SHV family similar to the analogous T_{266} in TEM (60). G_{238} is coupled with an essential catalytic residue Y_{105}, which further links with other catalytic functional residues: S_{70} and A_{237}, and R_{266}, a residue that known to stabilize the Ω-loop. This indicates that mutations on G_{238} would result in an alteration on protein catalytic function, as well as an increased flexibility of the protein, which strongly aligns with previous finding. Its linked mutation site R_{205} does not showing direct coupling with any catalytic residues. Instead, it is directly coupled with $R_{266,}$ which we mentioned as an Ω-loop stabilizer. Thus, it is not surprising that R_{205} substitution alone is never observed in nature (61), as it would not give significant evolutionary advantage to the protein.

Insights into unexplained functional sites of PDC-3

Unlike the extensively studied SHV-1, the functional roles of individual amino acids in PDC-3 remains largely unexplored. This gap in understanding serves as welcome challenge for interpreting the effects of mutations and the dynamic behaviour of PDC-3 from our results. Although several mutation hotspots, such as those on the Ω-loop (44), have been identified, very little is known about the specific contributions of individual amino acids to the functionality of PDC-3.

In PDC-3, mutations have primarily been reported in the Ω-loop. They enhance its flexibility to accommodate the bulky side chains of antibiotics, while deletions are more common in the R2-loop (39). DyNoPy detected five communities in total (Figure S4B) with all the four predominant Ω-loop mutations appeared in these communities. Community 3, 4 and 5 are explained in the supplementary information (Figure S7). Furthermore, DyNoPy also detected several previously unexplored Ω-loop residues.

G_{214}, a known mutation site in PDC-3, is the core residue in community 1. Another two essential mutation sites: E_{219} and Y_{221} also participate in this community, directly coupled with G_{214} (Figure 5A). G_{214} also has direct couplings with four other Ω-loop residues: A_{195}, A_{197}, G_{212}, and L_{216}. Previous results have demonstrated that substitutions of Glycine to Alanine or Arginine at 214 significantly destabilizes the Ω-loop (32). The strong correlation between G_{214} and these Ω-loop residues emphasizes the significant contribution of G_{214} towards the stability of the Ω-loop, which corroborates with previous results (32). Moreover, substitutions such as $G_{214}A$ and $G_{214}R$ and mutations on E_{219} and Y_{221} do not affect R2 loop flexibility, resulting in the smaller active site volume among variants (32) because none of the residues from the R2 loop are detected in this community offering plausible explanation to previously unexplained phenomenon.

G_{204} is the core residue of community 2, coupled with 73 other residues, most of which are distant from the catalytic site, suggesting plausible crucial role in overall protein stability like L_{162} in SHV-1 (Figure 5B). G_{204}, a newly emerged mutation site in the PDC family (62), is located on the short β-sheet β5a within the Ω-loop, near the hinge region between β8 and β9 just above the active site. The only known variant of G_{204} is PDC-466, which was derived from PDC-462 ($A_{89}V$, $Q_{120}K$, $V_{211}A$, $N_{320}S$), with an addition of $G_{204}D$ (62). Coupling of G_{204} to several catalytically important residues, including K_{67}, K_{315}, and T_{316} can suggest that mutations at this site can negatively impact catalytic power. This offers a plausible explanation of seeing fewer variants at this site and mutations at this site could have impact on hydrolysing capabilities of PDC variants. This should be confirmed by further experimental studies of variants of G_{204}. Unlike G_{214}, E_{219} and Y_{221} mutations which do not influence the dynamics of the R2 loop, substitutions on $V_{211,}$ a member of Ω-loop, has impact on dynamics of R2 loop because of its indirect couplings, through G_{204} to R2-loop residues (32). Two less critical substitution sites, H_{188} and V_{329}, were also observed in community 2.

Conclusions

DyNoPy combines information on residue-residue coevolution and protein dynamics captured from MD ensembles to detect conserved dynamic couplings. These couplings are model as a graph and network analysis facilitate the extraction of epistatic subnetworks and the assignment of roles to residues based on their important in the graph model. The choice of a relevant descriptor of protein dynamics has an impact on the ability to detect couplings that are involved in functional dynamics. Here we demonstrated how the choice of relevant global and local descriptors returns a higher number of effective couplings (greater than 0), and in turn leads to interpretable graph models and communities. In other systems, when multiple descriptors can be used to quantify functional conformational change, it is expected that they will differently modulate the effect of coevolution coupling. This will be reflected in a different structure of the associated graph models. This suggests the use of DyNoPy to generate comparative models in proteins with multiple functions associated to distinct dynamical changes.

Mutations of L_{162} and N_{136} have not yet emerged in SHV-1, but they are detected by DyNoPy as core residues for communities. These residues are strongly coupled with other functional important residues, which play critical roles in protein stability and catalytic activity. The identification of these couplings shows high consistency with previous studies and highlights the importance of L_{162} and N_{136} in SHV-1 functional dynamics. Given their central role in these communities, mutations in L_{162} and N_{136} can significantly alter protein function, suggesting their potential for future evolutionary changes. However, their strong relationships with these critical functional residues also suggest that mutation at these sites would need to be balanced to maintain protein function, providing an explanation for why such mutations have not yet emerged in SHV-1 (63). The ability of DyNoPy in detecting functionally important mutation sites was demonstrated via well-characterized mutation sites including R_{205} and G_{238} from SHV-1. Moreover, DyNoPy shows predictive ability on less-studied mutation sites as G_{156} and $A_{146,}$ by detecting critical residue couplings that coevolved with functional motions.

Based on the knowledge we have gained from analysis of SHV-1 functional protein dynamics we suggest that in PDC-3, mutations at G_{204} because of its significant conserved dynamic couplings can lead to new ESBL/IRBL clinical variants. We suggest that DyNoPy can be used as a

predictive tool to identify potential functional residues within this enzyme and guide future mutagenesis studies.

In summary, by integrating hidden evolutionary information with direct dynamic interactions, DyNoPy provides a powerful framework for identifying and analysing functional sites in proteins. The tool not only identifies key residues involved in local and global interactions, but also improves our ability to predict silent residues with previously unknown roles for future experimental testing. Our application of DyNoPy to broad-spectrum β-lactamases ESBLs and IRBLs demonstrates its potential to address key medical challenges such as antibiotic resistance by providing valid predictions on protein evolution.

Methodology

DyNoPy generates a graph representation of the protein structure that captures the couplings between amino acid residues contributing to the functional dynamics of the protein. Residues are represented as graph nodes, and conserved dynamic couplings are recorded as edges. Edge weights quantify the strength of these couplings. The model is built on two assumptions: residue pairs should have i) coevolved and their ii) time-dependent interactions correlate with a functional conformational change.

Therefore, edge weights (J_{ij}) for residue i and j are calculated as:

$$J_{ij} = \alpha\gamma_{ij} + \beta\rho_{ij} \quad where\ \alpha + \beta = 1 \tag{1}$$

where γ_{ij} is the scaled coevolution score and $\rho_{ij}$is the degree of correlation with the selected functional conformational change. α and β are weights assigned to γ_{ij} and ρ_{ij} that have a sum of one. The relative weight of the scaled coevolution score (α) is set to 0.5 in this study. When either of the assumptions listed above is not met, J_{ij} is set to zero.

Scaled coevolution scores

The occurrence of residue-residue coevolution can be estimated and quantified using probabilistic models of correlated mutations from deep multiple sequence alignments (MSA). DyNoPy supports generation of the MSA using the HH-Suite package (64) and calculation of scaled coevolution score (γ_{ij}) using CCMpred (65). First a pairwise residue coevolution matrix (***C***) is calculated, then these raw scores (c_{ij}) are divided by the matrix mean (Equation 2). All scores (s_{ij}) smaller than 1 are set to zero, and the remaining values are normalised by the maximum value (Equation 3):

$$s_{ij} = \frac{c_{ij}}{\langle C \rangle} \tag{2}$$

$$\gamma_{ij} = \begin{cases} 0, & s_{ij} < 1 \\ \frac{s_{ij}}{s_{max}}, & s_{ij} \geq 1 \end{cases} \quad (3)$$

Correlation with functional motions

The contribution of a residue pair to a selected functional motion is estimated by how much the change in interaction energy between the two residues over time is correlated with a collective variable (CV) describing the functional motion:

$$r_{ij} = cor(\varepsilon_{ij}(t), d(t)) \quad (4)$$

$$\rho_{ij} = \begin{cases} 0, & r_{ij} \leq 0.5 \\ r_{ij}, & r_{ij} > 0.5 \end{cases} \quad (5)$$

where $\varepsilon_{ij}(t)$ is the pairwise non-bonded interaction energy (see details in Supplementary Information) and $d(t)$ is the time-dependent value of the CV. Examples of CV and a discussion on the choice of the most relevant CV is presented in the results section. Correlation values smaller than 0.5 are set to 0.

Graph representation and analysis of conserved dynamic couplings

All pairwise conserved dynamic couplings (Equation 1) are collected into a square matrix $\boldsymbol{J}$. A graph is built from $\boldsymbol{J}$, using python-igraph v0.11 library (66). Nodes represent residues, and edges are drawn between nodes with positive J_{ij}. Edge weights are set to J_{ij}. The relative importance of the residues in this model of protein dynamics is calculated as eigenvector centrality of the nodes (67). Residues involved in extensive correlated dynamics with other highly connected residues have higher eigenvector centrality (EVC) scores. Groups of residues contributing to important collective motions are detected by community analysis of the graph structure. The Girvan-Newman algorithm is used to extract the community structure (68).

Adaptive Sampling Molecular Dynamics Simulations

MD simulation data was sourced from our previous studies (31, 32). To summarise, SHV-1 structural coordinates (PDB ID: 3N4I) were obtained from the Protein Data Bank and modified to the wild type by introducing the E104D mutation. Similarly, the PDC-3 structure was derived from PDC-1 (PDB ID: 4HEF) by a T105A substitution. Both enzymes were protonated at pH 7.0 using PropKa from the PlayMolecule platform (69). One disulfide bond between C_{77} and C_{123} was specified in SHV-1. Both structures were solvated with TIP3P water molecules in a periodic box with a box size of 10 Å. Ions were added to neutralize the overall charge of each system at 150mM KCl. Amber force field ff14SB was used for all MD simulations (70). After an initial minimisation of 1000 steps, both the enzymes were equilibrated for 5 ns in the NPT ensemble at 1 atmospheric pressure using the Berendsen barostat (71). The initial velocities for each simulation were sampled from the Boltzmann distribution at 300 K. Multiple Markov State Model (MSM)-based adaptively sampled simulations were performed for both proteins based on the ACEMD engine (72, 73). A canonical (NVT) ensemble with a Langevin thermostat (74) (damping coefficient of 0.1 ps^{-1}) and a hydrogen mass repartitioning scheme were employed to achieve time steps of 4 fs. For SHV-1, each trajectory spanned 60 ns with a time step of 0.1 ns, with a total of 593 trajectories. In the case of PDC-3, 100 trajectories were collected, each containing 3000 frames, lasting 300 ns. To manage the extensive datasets efficiently, trajectories were strategically stridden to ensure that a minimum of 30,000 frames were preserved for each system. The resulting trajectories are summarized in Supplementary Table S4.

Calculation and Selection of Collective Variables

Any vector that yields time-dependent information can be used as a collected variable (CV) to describe protein functional motions. The usefulness of DyNoPy is dependent on the choice of the CVs. To guide the selection of CVs, we selected 12 distinct features: radius of gyration (Rg), the first principal component (PC1), partial PC1 (PC1_partial), the first time-lagged independent component (TC1), partial TC1 (TC1_partial), global root mean square deviation (gRMSD), partial RMSD (pRMSD), dynamical RMSD (dRMSD), global solvent accessible surface area (gSASA), partial SASA (pSASA), active site pocket volume, and the number of hydrogen bonds (hbond). A description of the CVs, including the calculation methods and the residues used to calculate those

partial variables, is detailed in the supplementary information. CVs were subsequently used as input features for DyNoPy. The selection criteria for CVs were based on the number of residue pairs each CV could detect.

Data Availability

All files required to run the simulations (topology, coordinates, input), processed trajectories (xtc), corresponding coordinates (pdb), can be downloaded from the DOI https://doi.org/10.57760/sciencedb.15876 (PDC-3) and 10.5281/zenodo.13693144 (SHV-1). DyNoPy is available at https://github.com/alepandini/DyNoPy.

Acknowledgements

SCD was supported by Leverhulme Trust grant RPG-2017-222 awarded to AP and JAG. The authors would like to thank Arianna Fornili for insightful suggestions on the design of DyNoPy methodology.

References

1. P. D. Stenson *et al.*, The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677 (2017).
2. K. A. Matreyek *et al.*, Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874-882 (2018).
3. F. J. Poelwijk, V. Krishna, R. Ranganathan, The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Comput Biol* **12**, e1004771 (2016).
4. M. H. Høie, M. Cagiada, A. H. Beck Frederiksen, A. Stein, K. Lindorff-Larsen, Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep* **38**, 110207 (2022).
5. L. M. Blaabjerg *et al.*, Rapid protein stability prediction using deep learning representations. *Elife* **12** (2023).
6. A. S. Dunham, P. Beltrao, Exploring amino acid functions in a deep mutational landscape. *Mol Syst Biol* **17**, e10305 (2021).
7. T. A. Hopf *et al.*, Mutation effects predicted from sequence co-variation. *Nat Biotechnol* **35**, 128-135 (2017).
8. P. Radivojac *et al.*, A large-scale evaluation of computational protein function prediction. *Nat Methods* **10**, 221-227 (2013).
9. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
10. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
11. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
12. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072-1080 (2012).
13. A. Broom, K. Trainor, Z. Jacobi, E. M. Meiering, Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. *Structure* **28**, 717-726 e713 (2020).
14. D. Ding *et al.*, Protein design using structure-based residue preferences. *Nat Commun* **15**, 1639 (2024).
15. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A* **116**, 8852-8858 (2019).
16. W. P. Russ *et al.*, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440-445 (2020).
17. M. Cagiada *et al.*, Discovering functionally important sites in proteins. *Nat Commun* **14**, 4175 (2023).
18. S. J. Wodak *et al.*, Allostery in Its Many Disguises: From Theory to Applications. *Structure* **27**, 566-578 (2019).
19. P. Campitelli, T. Modi, S. Kumar, S. B. Ozkan, The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. *Annu Rev Biophys* **49**, 267-288 (2020).

20. C. H. M. Rodrigues, D. E. V. Pires, D. B. Ascher, DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci* **30**, 60-69 (2021).
21. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. *Nature* **450**, 964-972 (2007).
22. L. C. James, D. S. Tawfik, Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28**, 361-368 (2003).
23. D. Granata, L. Ponzoni, C. Micheletti, V. Carnevale, Patterns of coevolving amino acids unveil structural and dynamical domains. *Proc Natl Acad Sci U S A* **114**, E10612-E10621 (2017).
24. Y. Liu, I. Bahar, Sequence evolution correlates with structural dynamics. *Mol Biol Evol* **29**, 2253-2263 (2012).
25. D. J. Parente, J. C. Ray, L. Swint-Kruse, Amino acid positions subject to multiple coevolutionary constraints can be robustly identified by their eigenvector network centrality scores. *Proteins* **83**, 2293-2306 (2015).
26. L. Ponzoni, G. Polles, V. Carnevale, C. Micheletti, SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets. *Structure* **23**, 1516-1525 (2015).
27. L. Sutto, S. Marsili, A. Valencia, F. L. Gervasio, From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci U S A* **112**, 13567-13572 (2015).
28. R. A. Estabrook *et al.*, Statistical coevolution analysis and molecular dynamics: identification of amino acid pairs essential for catalysis. *Proc Natl Acad Sci U S A* **102**, 994-999 (2005).
29. Z. Chen, S. Rappert, J. Sun, A. P. Zeng, Integrating molecular dynamics and co-evolutionary analysis for reliable target prediction and deregulation of the allosteric inhibition of aspartokinase for amino acid production. *J Biotechnol* **154**, 248-254 (2011).
30. J. Wang, Y. Zhao, Y. Wang, J. Huang, Molecular dynamics simulations and statistical coupling analysis reveal functional coevolution network of oncogenic mutations in the CDKN2A-CDK6 complex. *FEBS Lett* **587**, 136-141 (2013).
31. E. Olehnovics *et al.*, The Role of Hydrophobic Nodes in the Dynamics of Class A beta-Lactamases. *Front Microbiol* **12**, 720991 (2021).
32. S. Chen *et al.*, Omega-Loop mutations control the dynamics of the active site by modulating a network of hydrogen bonds in PDC-3 beta-lactamase. *bioRxiv* 10.1101/2024.02.04.578824 (2024).
33. K. Poole, Resistance to beta-lactam antibiotics. *Cell Mol Life Sci* **61**, 2200-2223 (2004).
34. K. Bush, Past and Present Perspectives on beta-Lactamases. *Antimicrob Agents Chemother* **62** (2018).
35. K. Bush, Proliferation and significance of clinically relevant beta-lactamases. *Ann N Y Acad Sci* **1277**, 84-90 (2013).
36. A. Matagne, J. Lamotte-Brasseur, J. M. Frere, Catalytic properties of class A beta-lactamases: efficiency and diversity. *Biochem J* **330 (Pt 2)**, 581-598 (1998).
37. A. Philippon, G. Arlet, R. Labia, B. I. Iorga, Class C beta-Lactamases: Molecular Characteristics. *Clin Microbiol Rev* **35**, e0015021 (2022).
38. T. Palzkill, Structural and Mechanistic Basis for Extended-Spectrum Drug-Resistance

Mutations in Altering the Specificity of TEM, CTX-M, and KPC beta-lactamases. *Front Mol Biosci* **5**, 16 (2018).

39. G. A. Jacoby, AmpC beta-lactamases. *Clin Microbiol Rev* **22**, 161-182, Table of Contents (2009).
40. R. P. Ambler *et al.*, A standard numbering scheme for the class A beta-lactamases. *Biochem J* **276 (Pt 1)**, 269-270 (1991).
41. I. Galdadas *et al.*, Defining the architecture of KPC-2 Carbapenemase: identifying allosteric networks to fight antibiotics resistance. *Sci Rep* **8**, 12916 (2018).
42. A. Liakopoulos, D. Mevius, D. Ceccarelli, A Review of SHV Extended-Spectrum beta-Lactamases: Neglected Yet Ubiquitous. *Front Microbiol* **7**, 1374 (2016).
43. D. Pagan-Rodriguez *et al.*, Tazobactam inactivation of SHV-1 and the inhibitor-resistant Ser130 -->Gly SHV-1 beta-lactamase: insights into the mechanism of inhibition. *J Biol Chem* **279**, 19494-19501 (2004).
44. M. D. Barnes *et al.*, Deciphering the Evolution of Cephalosporin Resistance to Ceftolozane-Tazobactam in Pseudomonas aeruginosa. *mBio* **9** (2018).
45. C. F. A. Negre *et al.*, Eigenvector centrality for characterization of protein allosteric pathways. *Proc Natl Acad Sci U S A* **115**, E12201-E12208 (2018).
46. C. Pozzi *et al.*, Crystal Structure of the Pseudomonas aeruginosa BEL-1 Extended-Spectrum beta-Lactamase and Its Complexes with Moxalactam and Imipenem. *Antimicrob Agents Chemother* **60**, 7189-7199 (2016).
47. P. Bogaerts, C. Bauraing, A. Deplano, Y. Glupczynski, Emergence and dissemination of BEL-1-producing Pseudomonas aeruginosa isolates in Belgium. *Antimicrob Agents Chemother* **51**, 1584-1585 (2007).
48. L. Poirel *et al.*, BEL-2, an extended-spectrum beta-lactamase with increased activity toward expanded-spectrum cephalosporins in Pseudomonas aeruginosa. *Antimicrob Agents Chemother* **54**, 533-535 (2010).
49. F. Bo□s, J. Pleiss, Conserved water molecules stabilize the Omega-loop in class A beta-lactamases. *Antimicrob Agents Chemother* **52**, 1072-1079 (2008).
50. V. Agarwal, T. C. Yadav, A. Tiwari, P. Varadwaj, Detailed investigation of catalytically important residues of class A beta-lactamase. *J Biomol Struct Dyn* **41**, 2046-2073 (2023).
51. T. P. Cao *et al.*, Non-catalytic-Region Mutations Conferring Transition of Class A beta-Lactamases Into ESBLs. *Front Mol Biosci* **7**, 598998 (2020).
52. D. Parwana *et al.*, The Structural Role of N170 in Substrate-Assisted Deacylation in KPC-2 beta-Lactamase. *Angew Chem Int Ed Engl* **63**, e202317315 (2024).
53. I. Galdadas *et al.*, Allosteric communication in class A beta-lactamases occurs via cooperative coupling of loop dynamics. *Elife* **10** (2021).
54. S. Lu *et al.*, An active site loop toggles between conformations to control antibiotic hydrolysis and inhibition potency for CTX-M beta-lactamase drug-resistance enzymes. *Nat Commun* **13**, 6726 (2022).
55. J. E. Corkill, L. E. Cuevas, R. Q. Gurgel, J. Greensill, C. A. Hart, SHV-27, a novel cefotaxime-hydrolysing beta-lactamase, identified in Klebsiella pneumoniae isolates from a Brazilian hospital. *J Antimicrob Chemother* **47**, 463-465 (2001).
56. L. Poirel *et al.*, Emergence in Klebsiella pneumoniae of a chromosome-encoded SHV beta-lactamase that compromises the efficacy of imipenem. *Antimicrob Agents Chemother* **47**,

755-758 (2003).

57. N. Ben Achour, P. S. Mercuri, M. Ben Moussa, M. Galleni, O. Belhadj, Characterization of a novel extended-spectrum TEM-type beta-lactamase, TEM-164, in a clinical strain of Klebsiella pneumoniae in Tunisia. *Microb Drug Resist* **15**, 195-199 (2009).

58. M. H. Nicolas, V. Jarlier, N. Honore, A. Philippon, S. T. Cole, Molecular characterization of the gene encoding SHV-3 beta-lactamase responsible for transferable cefotaxime resistance in clinical isolates of Klebsiella pneumoniae. *Antimicrob Agents Chemother* **33**, 2096-2100 (1989).

59. C. R. Bethel *et al.*, Role of Asp104 in the SHV beta-lactamase. *Antimicrob Agents Chemother* **50**, 4124-4131 (2006).

60. A. P. Kuzin *et al.*, Structure of the SHV-1 beta-lactamase. *Biochemistry* **38**, 5720-5727 (1999).

61. S. Neubauer *et al.*, A Genotype-Phenotype Correlation Study of SHV beta-Lactamases Offers New Insight into SHV Resistance Profiles. *Antimicrob Agents Chemother* **64** (2020).

62. C. A. Colque *et al.*, Development of antibiotic resistance reveals diverse evolutionary pathways to face the complex and dynamic environment of a long-term treated patient. *bioRxiv* 10.1101/2021.05.14.444257 (2021).

63. M. Soskine, D. S. Tawfik, Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**, 572-582 (2010).

64. M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175 (2011).

65. S. Seemayer, M. Gruber, J. Soding, CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130 (2014).

66. G. Csárdi, T. Nepusz (2006) The igraph software package for complex network research.

67. M. E. J. Newman, Detecting community structure in networks. *The European Physical Journal B - Condensed Matter* **38**, 321-330 (2004).

68. M. E. Newman, Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys* **74**, 036104 (2006).

69. G. Martinez-Rosell, T. Giorgino, G. De Fabritiis, PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J Chem Inf Model* **57**, 1511-1516 (2017).

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

71. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684-3690 (1984).

72. S. Doerr, M. J. Harvey, F. Noe, G. De Fabritiis, HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J Chem Theory Comput* **12**, 1845-1852 (2016).

73. M. J. Harvey, G. Giupponi, G. D. Fabritiis, ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J Chem Theory Comput* **5**, 1632-1639 (2009).

74. R. L. Davidchack, R. Handel, M. V. Tretyakov, Langevin thermostat for rigid body dynamics. *J Chem Phys* **130**, 234101 (2009).

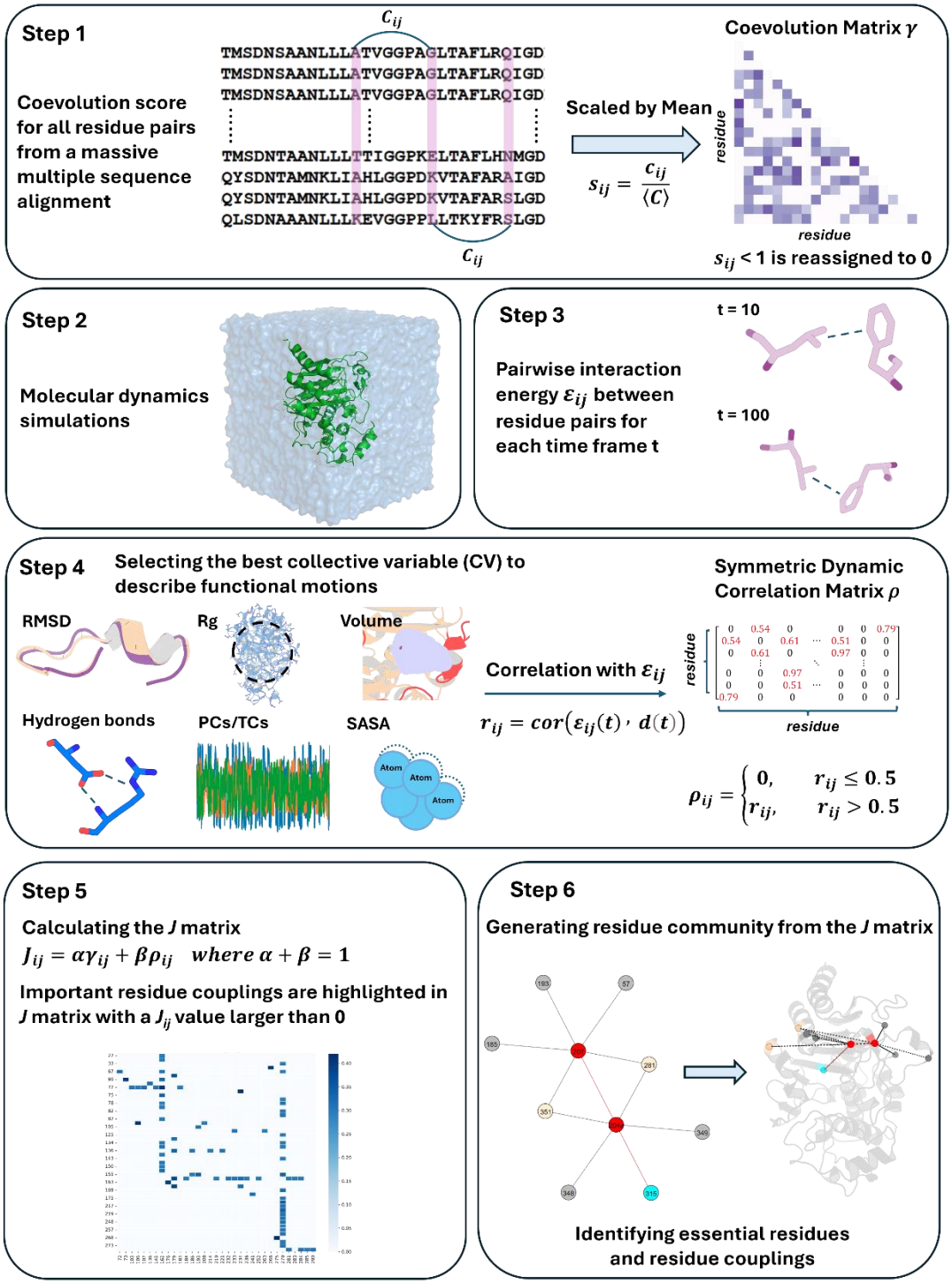

Figure 1 – Overview of DyNoPy Workflow

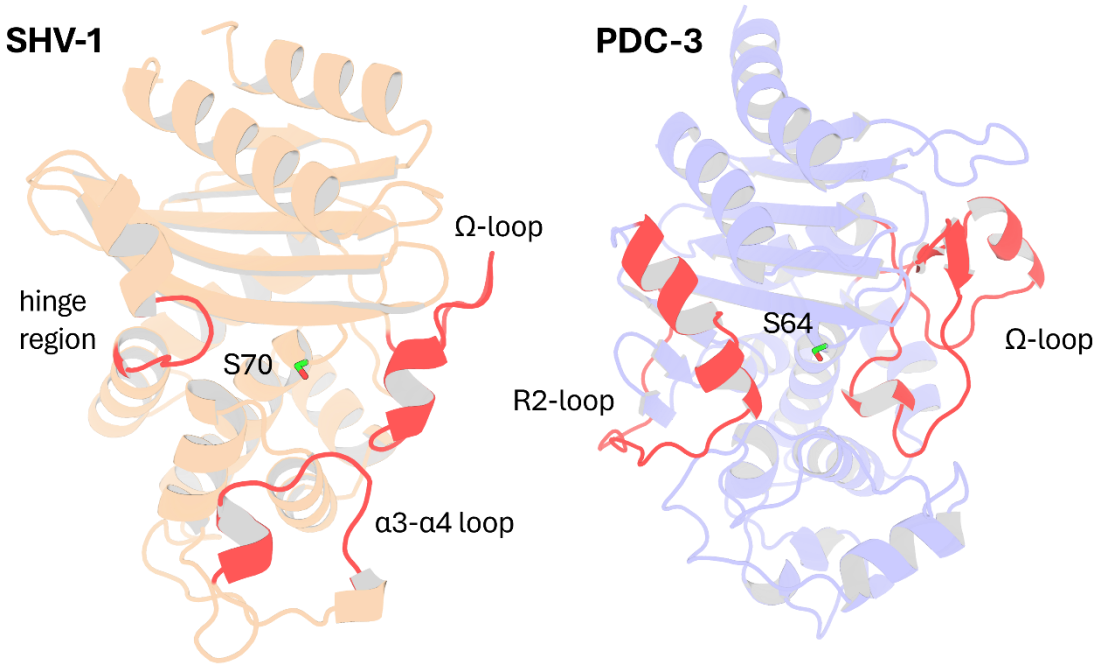

Figure 2 – Structural Comparison of SHV-1 (PDB ID: 3N4I) and PDC-3 (PDB ID: 4HEF) β-Lactamases. Catalytic serine S_{70} (SHV-1) and S_{64} (PDC3) are highlighted using stick representation. Important loops surrounding the active site are highlighted in red. In SHV-1, highlighted loops are the α3-α4 loop (residues 101-111), the Ω-loop (residues 164-179), and the hinge region (residues 213-218). In PDC-3, highlighted loops are the Ω-loop (residues 183-226) and the R2-loop (residues 280-310).

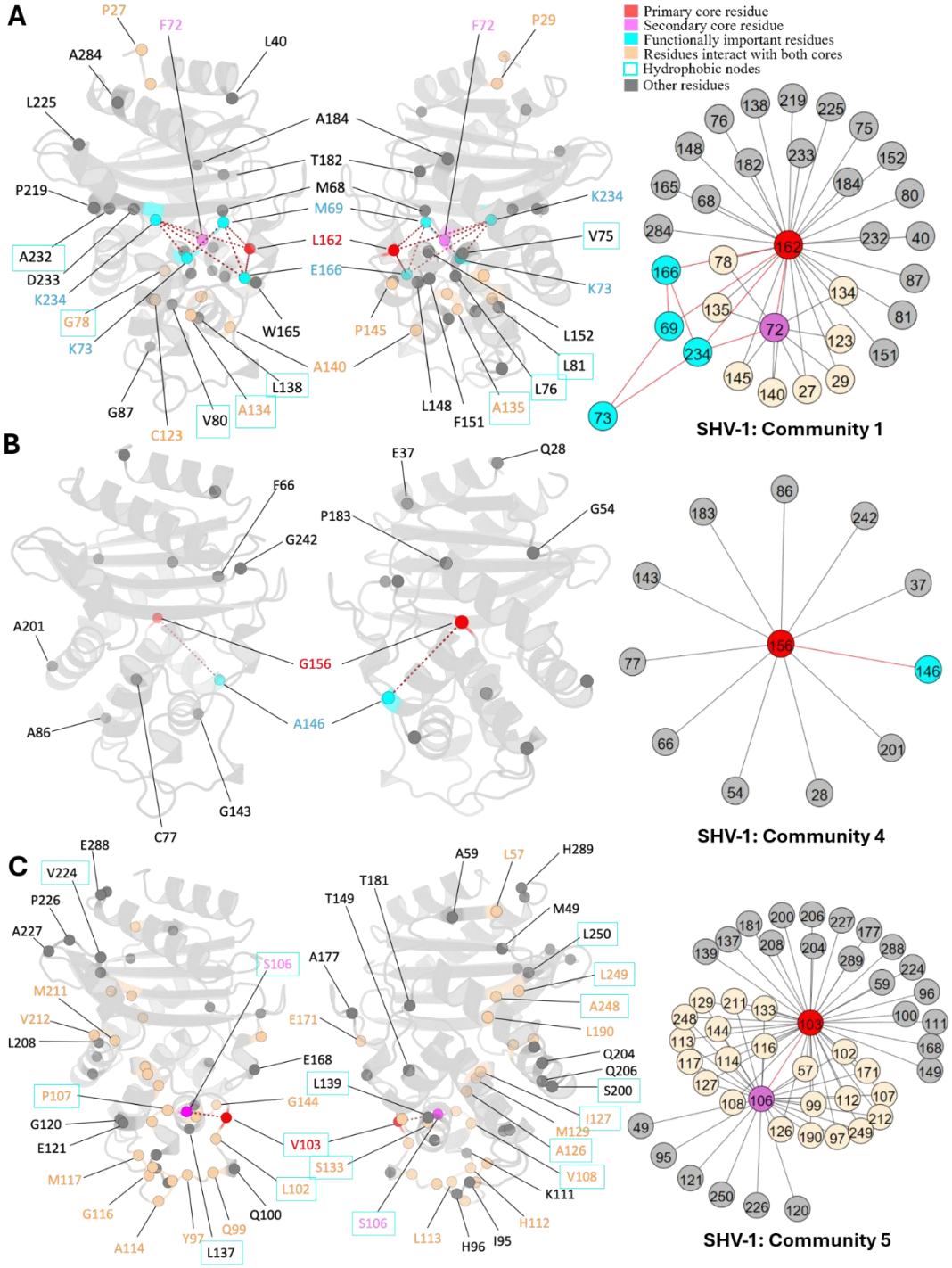

Figure 3 – Community 1, 4, and 5 of SHV-1 β-Lactamase. All the residues are depicted as spheres on the protein structure. The core residue for each community is highlighted in red, while purple is used to emphasize the secondary core residue. Residues that interact with both cores are coloured in light yellow. Functional important residues are marked in cyan. Hydrophobic nodes are enclosed with cyan boxes. A. Community 1 of SHV-1, comprising 33 residues with L162 being the primary core residue. B. Community 4 of SHV-1, containing 12 residues and is centred by G156. C. Community 5 of SHV-1, embracing 48 residues and showing a strong correlation between V103 and S106.

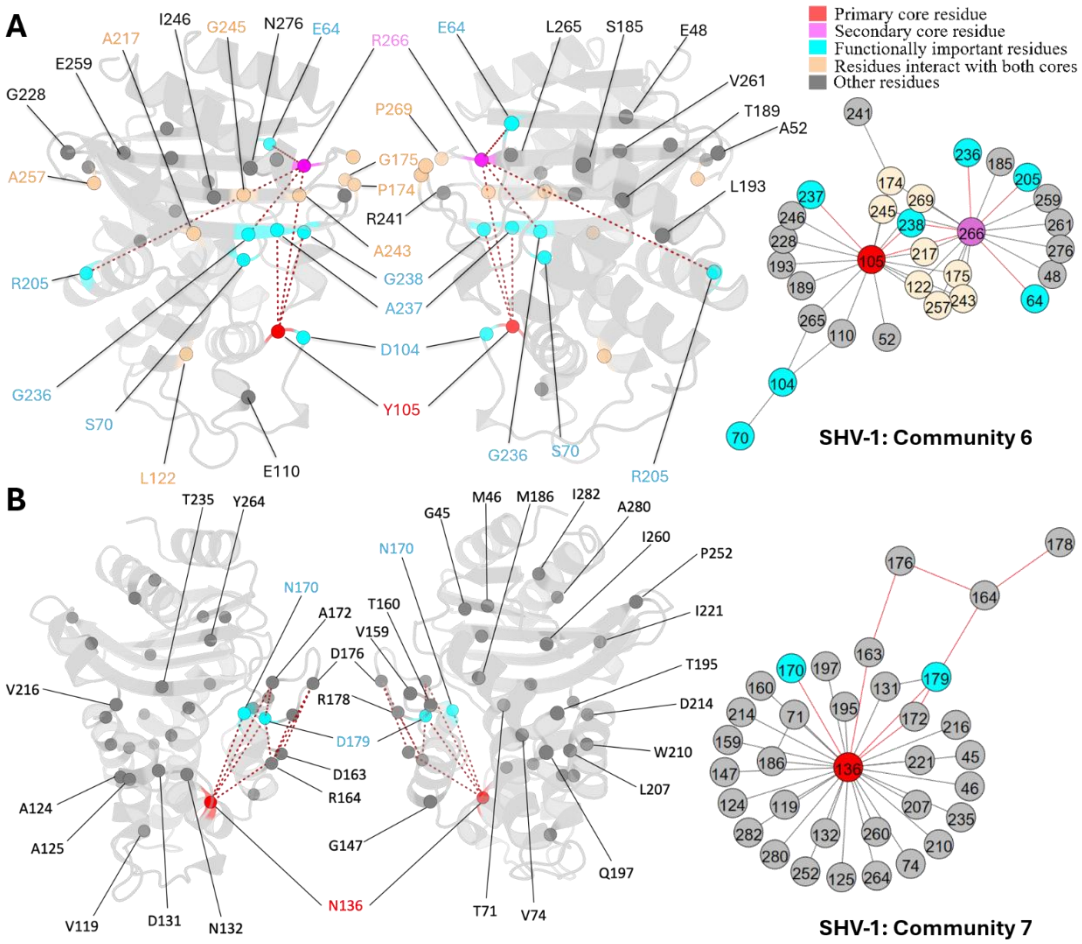

Figure 4 – Community 6 and 7 of SHV-1 β-Lactamase. All the residues are depicted as spheres on the protein structure. The core residue for each community is highlighted in red, while purple is used to emphasize the secondary core residue. Residues that interact with both cores are coloured in light yellow. Functional important residues are marked in cyan. A. Community 6 of SHV-1, comprising 30 residues with Y_{105} being the primary core residue. B. Community 7 of SHV-1, containing 34 residues and is centred by N_{136}.

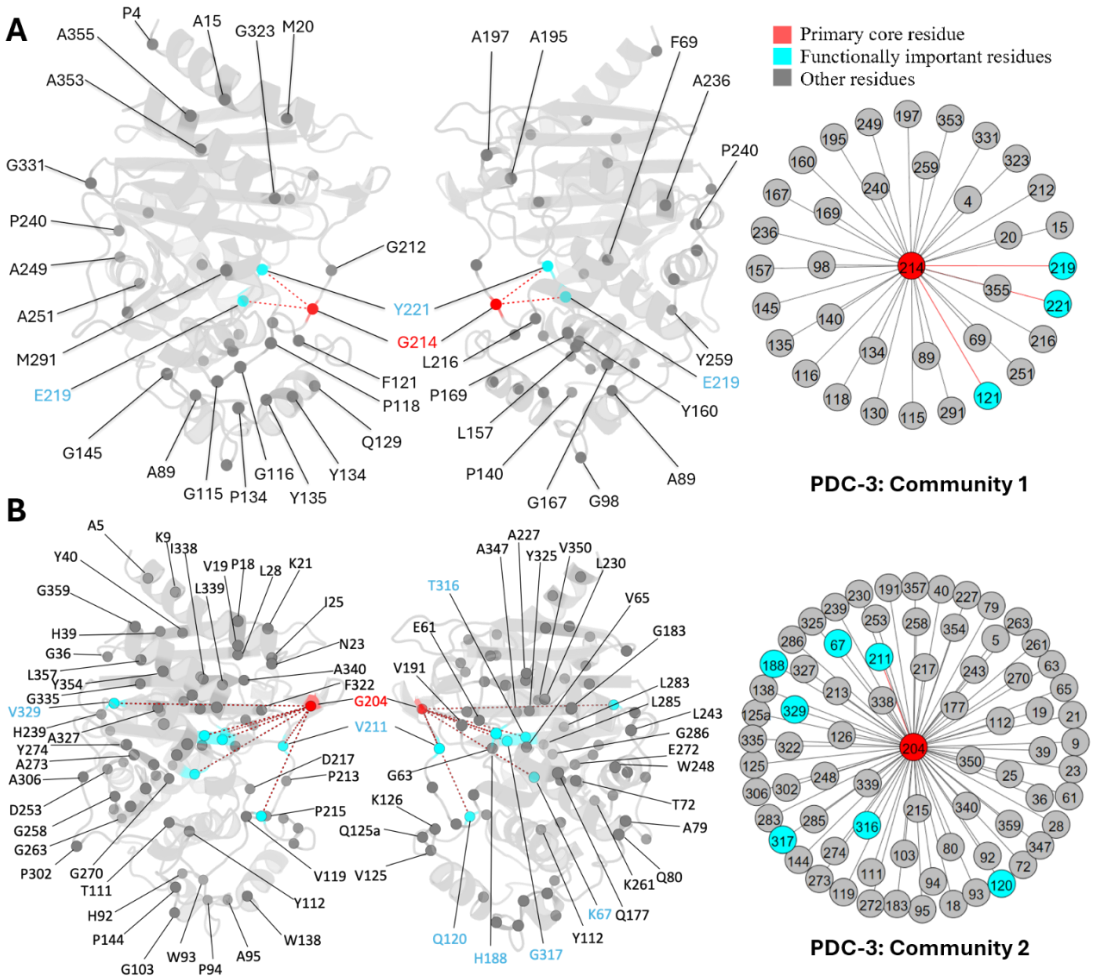

Figure 5 - Community 1 and 2 of PDC-3 β-Lactamase. All the residues are depicted as spheres on the protein structure. The core residue for each community is highlighted in red. Functional important residues are marked in cyan. A. Community 1 of PDC-3, comprising 36 residues with G_{214} being the primary core residue. B. Community 2 of PDC-3, containing 74 residues and is centred by G_{204}.