Quantitative structural interpretation of protein cross-links

Isaac Filella-Merce[1,2], Benjamin Bardiaux[1], Michael Nilges[1], and Guillaume Bouvier[1,3,*]

[1]Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, Institut Pasteur, CNRS UMR3528, C3BI, USR3756 Paris, France
[2]Faculty of Health and Life Sciences, University Pompeu Fabra, Carrer del Doctor Aiguader 80, Barcelona, 08003, Spain
[3]Lead Contact
[*]Correspondence: guillaume.bouvier@pasteur.fr

ABSTRACT

Chemical cross-linking, combined with mass spectrometry analysis, is a key source of information for characterizing the structure of large protein assemblies, in the context of molecular modeling. In most approaches, the interpretation is limited to simple spatial restraints, neglecting the physico-chemical interactions between the cross-linker and the protein and of flexibility. Here we present a method, named NRGXL (New Realistic Grid for Cross-Links), which models the flexibility of the cross-linker and the linked side chains, by explicitly sampling many conformations. Also, the method can efficiently deal with overall protein dynamics. This method creates a physical model of the cross-linker and associated energy. A classifier based on it outperforms others, based on Euclidean distance or solvent accessible distance and its efficiency makes it usable for validating 3D models from cross-linking data. NRGXL is freely available as a web server at: https://nrgxl.pasteur.fr.

Keywords

NRGXL, Cross-links, sampling, modeling, protein complexes, restraints, Binary Classification study.

Introduction

Macromolecular assemblies play a vital role in cellular processes. Knowing their structure is essential to get a detailed understanding of their function. Most of the time, a single standard structure determination method cannot be used for these assemblies due to their size, their flexible, or their transient nature, and strategies combining structural data from different sources (Ward et al., 2013; Lasker et al., 2010) are increasingly being employed (Robinson et al., 2015; Lasker et al., 2012; Alber et al., 2009; Chen et al., 2010). Due to its relative experimental simplicity, chemical cross-linking coupled with Mass Spectrometry (XL-MS) (Leitner et al., 2016; Holding, 2015; Leitnera et al., 2010; Rappsilber, 2011; Bullock et al., 2018) plays a key role in integrative structural biology.

XL-MS uses chemical cross-linking agents (cross-linkers) together with mass spectrometry and database searching in order to identify pairs of residues covalently bonded by the cross-linker. If we know that two residues of a given complex are cross-linked, one can deduce that the distance between them cannot be longer than the length of the cross-linker. This information can be used in two ways: as a criterion if a particular conformation is compatible with the chemical cross-link, or as a restraint to guide molecular modeling. In this work, we mostly focus on the first aspect, but we will discuss how the method can be efficiently used directly in molecular modeling.

Several strategies have been proposed for the structural interpretation of cross-links. The simplest is to use the distance between the C_α carbons of the cross-linked residues as a Euclidean metric (Robinson et al., 2015; Chen et al., 2010; Ferber et al., 2016) (straight line, Figure 2a). For example, Brodie *et al.* (Brodie et al., 2017) used cross-linking data to model protein structures by restrained molecular dynamics simulations. The crucial point of this approach was to use short-distance cross-linkers to be able to approximate them as an Euclidean distance restraints during the modeling. Although this approximation is suitable for a short cross-linker, this is a rough approximation for longer ones, since it neglects the fact that the bulky cross-linkers cannot physically overlap with the protein atoms. This criterion classifies two residues with a Euclidean distance lower than the cross-linker length, but deeply buried in the protein, as cross-linkable, even though there is physically no place to put the cross-linker, resulting in a false positive prediction.

Computing the shortest solvent-accessible surface distance (SASD) between the two residues through free space (Figure 2a) circumvents this limitation (Degiacomi et al., 2017; Kahraman et al., 2011; Bullock et al., 2016; Ferrari

et al., 2019b). However, this method has limitations. For example, for two residues located in a narrow protein cavity, the SASD between the residues may be shorter than the cross-linker length, but the cavity may be too small to accommodate the cross-link.

Also, XL-MS data come from an ensemble of conformations. In particular, lysine residues, which are typically cross-linked, are often very flexible, and this flexibility needs to be accounted for when comparing a particular molecular conformation to cross-linking data. Degiacomi et al. have shown (Degiacomi et al., 2017) that even a simple geometric representation of side-chain flexibility (considering multiple positions of side chain head groups on a sphere around the C_β carbons) increases the accuracy of cross-link assessment.

The approach we present here is designed to address all these limitations together, by taking into account the physical energy of the cross-linker bound to the protein, and the flexibility of the cross-linked side chains and the linker. Our method discards the restrictive idea of a simple geometric interpretation and focuses on the physico-chemical interactions between the cross-linker and the protein, by explicitly modeling conformations of the cross-linker and the side chains that do not overlap with the protein matrix. Flexibility is taken into account by sampling multiple conformations of the side chains and the cross-linker. This is illustrated in Figure 2b. A Constrained Markov Chain algorithm, followed by local energy minimization, efficiently sample cross-linker conformations. Furthermore, the method can take the flexibility of the protein matrix into account by estimating protein motions with the anisotropic elastic network model (ANM).

The structural model used for describing the association of cross-linkers and the protein matrix is more realistic than purely geometric methods. Consequently, when using the conformational energy as a classifier, it outperforms all other proposed methods. The efficiency of the method makes it usable in the context of molecular modeling of protein assemblies from cross-linking data.

Results

NRGXL: cross-link sampling approach

The approach, called NRGXL (New Realistic Grid for Cross-Links), assesses the propensity of a cross-link to be formed between two residues in a given structure by going beyond a geometric approach and considering the physical interaction between the cross-linker and the protein. The method is based on a sampling process which, given two residues, generates multiple cross-link conformations between them (Figure 2b) in the context of the protein matrix. These conformations differ in the position of the atoms of the two linked side chains and of the linker itself (Figures 2b). In this way, the flexibility of both the side chains and the linker are explicitly modeled, without making geometric oversimplifications.

NRGXL employs the Constrained Markov Chain algorithm on a 3D grid to sample the space that is accessible to the side chains and the linker (Figure 1). Starting from the C_β atom of the first residue, the atoms of the first side chain, the linker, and the second side chain are iteratively placed on a neighbouring grid point with a probability depending on the occupancy of the grid points (i.e., if the grid point is close to an atom of the protein), under the constraint that the constructed path has to connect the two C_β atoms of the cross-linked residues. This is repeated several times to sample different possible conformations that are compatible with the length of the cross-linker and the surrounding protein. These initial conformations on the 3D grid are then minimized in a standard refinement force field, allowing the atoms to move off-grid, and the resulting conformational energy is stored. The result is an ensemble of conformations, each associated with an energy value. The conformation of minimum energy is selected as the representative cross-linker structure between the given residues, and its energy is used as a classifier. The method is described in more detail in the METHOD DETAILS section.

To evaluate NRGXL and compare it to the other available methods, we used the entries from the cross-link database (XLdb) (Kahraman et al., 2013) (see XLdb), a database of pairs of chemical cross-links and the independently solved corresponding X-ray crystal structure. For each entry in the XLdb, we determined the set of compatible lysine pairs as all pairs with a Euclidean distance shorter than the length of the cross-linker plus the two side chains (28.42 Å when using Bissulfosuccinimidyl suberate (BS3) as the cross-linker and two lysines as the cross-linked residues), see Binary Classification study.

Cross-link ensembles in a *static* and a *flexible model*

The propensity of forming a cross-link is not only influenced by local flexibility but also by overall protein motion. We compared the influence of a simple model of backbone flexibility, the Anisotropic Network Model (ANM) (Atilgan et al., 2001; Bakan et al., 2011), on the performance of NRGXL. For this, we first used each deposited structure in the XLdb in a *static model*. We note that the *static model* still contains full flexibility for the cross-linked side chains and the linker. We then sampled protein conformations with the ANM (see Anisotropic Network Model), generating 10 conformations from each deposited XLdb structure (*flexible model*). In this way, we included flexibility not only of the side chains but also of the backbone of the whole complex in a simple and efficient way.

In the *static model*, 24 systems had at least one compatible lysine pair satisfying the above criterion, with a total of 236 compatible lysine pairs with an experimentally detected cross-link associated (EyXL) and 7613 compatible lysine pairs without (EnXL). In contrast, in the *flexible model*, we could include 27 systems in our analysis, accounting for 273 compatible lysine pairs with an associated experimentally detected cross-link, and 13089 compatible lysine pairs without. The nine extra conformations can easily explain the differ-

ence in number of systems and number of compatible lysine pairs between the *static* and *flexible models* per database entry in the latter. The sampling induces differences also in the distances in lysine pairs. If a particular distance is shorter in at least one of the nine additional conformations than in the static conformation, the corresponding residue pair is included in the analysis.

For each compatible lysine pair detected in the *flexible* or the *static model*, an ensemble of cross-linker conformations was computed with NRGXL. The maximum number of cross-link conformations generated for each lysine pair is a user-defined parameter for NRGXL (set to 20 in the present analysis). In some cases, the algorithm finds less than the desired number of conformations, in essence, if the lysine pair is too buried inside the protein matrix (see Constrained Markov Chain algorithm).

Binary Classification: study of cross-linkability

To compare the ability of different criteria to predict the propensity of a lysine pair to be cross-linked, we conducted a Binary Classification study (see Binary Classification study), for the compatible lysine pairs of both the *static* and *flexible models*. Due to the unbalance between the number of EyXL and EnXL, we randomly selected even populations of EyXL and EnXL. In this classification study, four different features were analyzed, the Euclidean distance (EucDist) between the C_β atoms of the two lysines; SASD between the C_β atoms of the two lysines, computed with the SciPy library (Jones et al., 2001); SASD, between the N_ζ atoms of the two lysines, computed with DynamXL (DXL) (Degiacomi et al., 2017); and the energy computed with NRGXL.

For a Binary Classification, one has to set a threshold for a quantitative feature (here, a distance or an energy value), which is then used to classify the elements. In the present case, we classified compatible lysine pairs into two sets according to an upper threshold: the set predicted to be in cross-linkable conformations (CyXL) when the considered value is lower than the threshold and, the set predicted not to be in cross-linkable conformations (CnXL) for values higher than the threshold.

As one would expect, the distributions of distances for EucDist for the two cases EyXL and EnXL (compatible distance with and without cross-link in the database, respectively) are different, with the mode of the distribution shifted towards larger values for EnXL (Figure 3). We observed the same behaviour for SASD and energies (data not shown).

A standard measure to evaluate the quality of a feature as a classifier is the Area Under the Curve (AUC) of its Receiver Operating Characteristic curve (ROC curve) (Hand, 2009). Generally, the higher the AUC of a given feature, the better it performs the task of Binary Classification. We computed the ROC plots by varying the feature thresholds between a minimum and a maximum value and calculating the false positive and true positive rates. Using 100 randomly selected EyXL/EnXL balanced populations, we generated averaged ROC curves for each feature and model (Figure 4). Figure 5 shows that for both the *static* and the *flexible model*, the ROC curve with the highest AUC is the one representing the energy computed with NRGXL, with AUC values of 0.773 ± 0.016 and 0.786 ± 0.014, respectively; the difference between NRGXL and the next best method, is even more pronounced in the *flexible model*. Interestingly, Degiacomi *et al.* (Degiacomi et al., 2017) also found that the inclusion of different conformations like in the *flexible model* can be important when explaining the experimental data. Moreover, we also computed the SASD using the XWalk software (Kahraman et al., 2011), one of the most established programs to compute cross-link distances based on an SASD evaluation. We found an AUC of 0.75 ± 0.02 for the static model, which is sensibly better than our SASD implementation but comparable with the AUC obtained using DynamXL (see Figure 5).

To assess the effect of the size of the protein on the predictive ability of the classifiers, we split the database into three groups based on the size of the protein (0-2000, 2000-4000 and 4000-6000 residues). We saw that the gain of the energy classifier relative to the EucDist classifier and the SASD is even higher for larger systems (see Figure S1).

In some cases, the AUC of a ROC plot is not the best performance indicator for predictors. For example, the AUC can give misleading results (Hand, 2009) if ROC curves cross each other or have different misclassification costs (false-positive (FP) and false-negative (FN) relevance, respectively) for their respective classifiers. In our case, the ROC curves for NRGXL never cross the other ROC curves (Figure 4). Also, during the whole study, we considered equal misclassification costs (FP is considered equally bad as FN for all the classifiers). In consequence, AUC is a good indicator of classification performance in the present case.

Another good indicator is the accuracy, defined as the fraction of correctly classified cases (true positives and true negatives), in our case, the fraction of compatible lysine pairs that were correctly classified. Figure 6 shows the average accuracy depending on the energy threshold in both the *static* and *flexible models* over 100 randomly selected EyXL/EnXL balanced populations. 50 rounds of independent training and validation were set up by splitting the database into two sets. For each training, equally balanced populations were picked up 100 times to set the optimal threshold, by maximizing the accuracy. Then, this threshold was used to compute the accuracy and precision on the corresponding validation set. This procedure was simultaneously applied to all features. In the *static model*, the highest value of accuracy reached during the training was 0.75 ± 0.02, for an energy of 1.35 ± 12.19 kcal $\cdot$ mol^{-1} corresponding to an accuracy of 0.74 ± 0.02 during validation (Figure 6a and Table 1). In the *flexible model*, the highest accuracy of 0.76 ± 0.02 was attained during training at an energy threshold of -8.83 ± 5.91 kcal $\cdot$ mol^{-1}, corresponding to an accuracy of 0.75 ± 0.02 in validation (Figure 6b and Table 1). Table 1 summarizes the highest

values of accuracy from the training, their corresponding optimal thresholds and the resulting accuracies/precisions from the validation for all features in the two studied models. The feature with highest accuracy and precision in the validation subset (for both models) is the energy computed with NRGXL. Interestingly, the optimal threshold for the EucDist is much shorter than the straightforward C_β distance corresponding to the theoretical maximum distance where a cross-link can be formed between lysine pairs (28.42 Å, cross-linker length plus twice the lysine length). As an extension of Table 1, Table S1 includes three additional statistical measures obtained during the validation process, sensitivity, specificity and negative predictive value. Likewise accuracy and precision maximum values of those new outcomes were obtained when using NRGXL. Table S1 also presents an alternative optimal threshold selection method based on ROC curves (closest-to-(0,1) (Unal, 2017)) which results in similar statistical measures.

To ensure that maximum accuracies used to select our optimal thresholds do not overlap with the distribution of the accuracy of a random classifier (black dotted lines in Figure 6, we computed a p-value for each one of them (Figure 6) (more details of how this p-value was calculated in Binary Classification study: treatment of random errors.) In both models, and for all four features, the highest values of accuracy had p-values close to 0.

Moreover, in the case of energy, the values of accuracy with p-values over 0.05, which are not significant, are concentrated in the lower and upper limits of the energy range of values (Figure 6). This result is because taking a left or right limit feature values as thresholds implies to classify almost all compatible lysine pairs in one of the two categories CyXL or CnXL. When picking a left limit threshold, we are classifying the vast majority of compatible lysine pairs as CnXL, and when picking a right limit threshold, we are classifying almost all compatible lysine pairs as CyXL. This "boundary" behavior is independent of the feature studied.

Finally, the outcomes of one Binary Classification study using all available compatible lysine pairs per feature and model is shown in Figure 7. Here, we took as cut-offs the earlier calculated optimal thresholds. It is important to keep in mind that the total number of compatible lysine pairs ($TP+TN+FP+FN$) remains constant between features of the same model. An increased number of TN compared to FP explains the major improvement of the energetic approach relative to the geometrical approaches (DynamXL, SASD and EucDist). This can be quantified by the specificity, which is defined as the fraction of EnXL correctly classified as true negative (specificity $= \frac{TN}{TN+FP}$). This tendency is obvious for the *static model* where the number of TN is constantly increasing from the Euclidean based approach to the SASD, DynamXL and finally the energetic approach which reaches a specificity of 0.71. See Binary Classification study for more details on the generation of these outcomes.

Efficiency

The results of the efficiency test are presented in Table.2. As expected, no dependency is found in between the time needed for the sampling of an ensemble of cross-link conformations per residue pair and the size of the complex. Nonetheless, we can see a dependency of the overall time (sum of the sampling time and the preparation time) required for generating an ensemble and the size of the complex. Whereas the sampling time is the time needed for the Markov Chain algorithm to generate the ensemble in the context of a small local grid encircling the targeting residues, the preparation time is the time needed to, through a whole complex grid, compute the smaller sampling grid (more details in Sampling approach). Hence, the preparation time is the one carrying the latter mentioned dependency observed in the overall time.

Discussion

NRGXL, the approach to assess the propensity of a residue pair to be cross-linked departs from the traditional, geometrical cross-link interpretation to focus on the physico-chemical interactions between the cross-linker and the protein. In terms of XL-MS data interpretation, the two main improvements of our sampling method are: (1) a more realistic structural model of the cross-linker (2) and an efficient and accurate model to treat the flexibility of the XL, the side chains, and the whole complex.

The Binary Classification study, where we analyzed the performance of different XL statistics as predictors of cross-linkability, clearly showed that spatial statistics used by previous studies such as Euclidean distance (Robinson et al., 2015; Chen et al., 2010; Ferber et al., 2016) or SASD (Degiacomi et al., 2017; Kahraman et al., 2011) perform less well than the energy computed by NRGXL. This indicates that the explicit cross-linker structures, including their pyshico-chemical interactions with the protein, are in better agreement with XL-MS data. This emphasizes the importance of taking into account the physico-chemical interactions between the cross-linker and the protein when interpreting XL-MS data.

The optimal thresholds determined in this study differ from what one could naively expect and should be the ones used when assessing a structure. The fact that the EucDist optimal threshold was much lower than the straightforward C_β EucDist (which would be the usual distance restraint used when modeling) highlights the importance of carefully evaluating the influence of a threshold in binary classification. Further studies using these thresholds also as values for restraints during modeling should be performed. We note that the optimal EucDist value found in this study is not only close to the value used in our modeling of the Pol III RNA polymerase (Ferber et al., 2016) but also in the range of a more recently proposed threshold by Ferrari *et al*. (Ferrari et al., 2019a) which was measured by an independent approach in a different database.

Our method outperforms the geometrical approaches especially by classifying better the FP to TN (Figure 7).

NRGXL overcomes the limitations of EucDist and SASD: one can imagine two residues in a small cavity with a distance compatible with the cross-linker length and without an experimentally detected cross-link between them (EnXL). If we determine the EucDist and SASD between these residues, they will be shorter than the cross-linker length. Hence, both EucDist and SASD will classify this pair of residues as CyXL and consequently generate an FP. In contrast, the energy of an XL conformation generated between them will be high if there is not enough space to properly accommodate the cross-linker in the cavity, leading to classification as CnXL, corresponding to a TN. Therefore the physical model developed here better describes the steric hindrance of the cross-linker that is not taken into account either in the EucDist nor in the SASD approach. For the number of TP, our predictor does not show significant improvement compared to the geometrical approach. This tendency is in part due to the incompleteness of XL-MS data since only a subset of cross-linkable lysines is detected as being linked, thus inducing cross-links labeled as FP which should be labeled as TP.

Due to its efficiency and accuracy, NRGXL can be used as a more realistic description of cross-linking restraints during modeling. However, the computational cost of NRGXL is evidently higher than calculating an Euclidean distances. Therefore, NRGXL could be used as a post-modeling filter, as it has already been suggested for SASD (Merkley et al., 2014), to assess the quality of the models resulting from a classical modeling approach using Euclidean distance restraints, or be used as part of a modeling strategy that uses a full calculation of realistic distances relatively rarely (e.g., (Ferber et al., 2016)). The efficiency is achieved without neglecting the importance of flexibility. Residue side chains experience rapid motions in solution (Henzler-Wildman and Kern, 2007), and XL-MS data represents an ensemble of different conformations. Hence, sampling several conformations for a given cross-link allows our method to deal with multiple side chain orientations and therefore incorporate side chain flexibility. The importance of treating also the overall flexibility is evidenced by the comparison between the *static* and *flexible models*. While in the *static model* we only considered side chain flexibility, in the *flexible model* we explicitly generated multiple conformations of the complete system with an anisotropic network model. This improves the performance not only of the approaches based on EucDist and SASD but also of NRGXL. The accuracy and efficiency of our method make it the method of choice to be used during the modeling of macro-molecular complexes.

Acknowledgements

We would like to thank R. Pellarin and B. Worley for helpful discussions and support. I.F.M. acknowledges support from the Erasmus+ framework. This work was supported by the European Research Council (MN: FP7-IDEAS- ERC 294809).

Author contributions statement

I.F.M., M.N. and G.B. worked on the conceptualization of the project. M.N. and G.B. supervised the project. The software development was conducted by I.F.M. and G.B. with a contribution by B.B. regarding the implementation of CNS. I.F.M. and G.B. analyzed the data. I.F.M., G.B., B.B., and M.N. wrote the manuscript.

Declaration of Interests

The authors declare no conflicts of interests.

References

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2009). Determining the architectures of macromolecular assemblies. *Nature*, 450:683–694.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal*, 80:505–515.

Bakan, A., Meireles, L. M., and Bahar, I. (2011). ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, 27:1575–1577.

Brodie, N. I., Popov, K. I., Petrotchenko, E. V., Dokholyan, N. V., and Borchers, C. H. (2017). Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Science Advances*, 3(7):e1700479.

Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nature Protocols*, 2:2728–2733.

Brunger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998). Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Cryst*, 54:905–921.

Bullock, J. M. A., Schwab, J., Thalassinos, K., and Topf, M. (2016). The Importance of Non-accessible Crosslinks and Solvent Accessible Surface Distance in Modeling Proteins with Restraints From Crosslinking Mass Spectrometry. *Mol Cell Proteomics*, 15(7):2491–500.

Bullock, J. M. A., Sen, N., Thalassinos, K., and Topf, M. (2018). Modeling Protein Complexes Using Restraints from Crosslinking Mass Spectrometry. *Structure - Cell Press*, 26:1015–1024.

Chen, Z. A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J., Nilges, M., Cramer, P., and Rappsilber, J. (2010). Architecture of the RNA polymerase II–TFIIF complex revealed by cross-linking and mass spectrometry. *The EMBO Journal*, 29:717–26.

Degiacomi, M. T., Schmidt, C., Baldwin, A. J., and L.P., J. (2017). Accommodating Protein Dynamics in the Mod-

eling of Chemical Crosslinks. *Structure - Cell Press*, 25:1751–1757.

Farabella, I., Vasishtan, D., Joseph, A. P., Pandurangan, A. P., Sahota, H., and Topf, M. (2015). TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J. Appl. Cryst*, 48:1314–1323.

Ferber, M., Kosinski, J., Ori, A., Rashid, U. J., Moreno-Morcillo, M., Simon, B., Bouvier, G., Batista, P. R., Müller, C. W., Beck, M., and Nilges, M. (2016). Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nature Methods*, 13:512–520.

Ferrari, A., Gozzo, F., and Martínez, L. (2019a). Statistical force-field for structural modeling using chemical crosslinking/mass spectrometry distance constraints. *Bioinformatics*.

Ferrari, A. J. R., Clasen, M. A., Kurt, L., Carvalho, P. C., Gozzo, F. C., and Martínez, L. (2019b). TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics*.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450:964–972.

Holding, A. N. (2015). XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods*, 89:54–63.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Version 0.18.1] [Online; accessed 8/5/2018].

Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R., and Malmström, L. (2013). Cross-Link Guided Molecular Modeling with ROSETTA. *PLoS One*, 8(9):e73411.

Kahraman, A., Malmström, L., and Aebersold, R. (2011). Xwalk: computing and visualizing distances in crosslinking experiments. *Bioinformatics*, 27:2163–2164.

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proccedings of the National Academy of Sciences*, 109:1380–1387.

Lasker, K., Phillips, J. L., Russel, D., Velázquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A., and Sali, A. (2010). Integrative Structure Modeling of Macromolecular Assemblies from Proteomics Data. *Mol Cell Proteomics*, 9:1689–1702.

Leitner, A., Faini, M., Stengel, F., and Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences*, 41:20–32.

Leitnera, A., Walzthoenia, T., Kahramana, A., Herzoga, F., Rinnera, O., Becka, M., and Aebersolda, R. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular and Cellular Proteomics*, 9:1634–49.

MacKerell, A. D., Bashford, D., Bellot, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wirkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102(18):3586–3616.

Merkley, E. D., Rysavy, S., Kahraman, A., Hafen, R. P., Daggett, V., and Adkins, J. N. (2014). Distance restraints from crosslinking mass spectrometry: {Mining} a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Science*, 23(6):747–759.

Rappsilber, J. (2011). The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of Structural Biology*, 173:530–540.

Robinson, P. J., Trnka, M. J., Pellarin, R., Greenberg, C. H., Bushnell, D. A., Davis, R., Burlingame, A. L., Sali, A., and Kornberg, R. D. (2015). Molecular architecture of the yeast Mediator complex. *eLife*.

Schüttelkopf, A. W. and van Aalten, D. M. F. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr*, 60:1355–1363.

Unal, I. (May 2017). Defining an optimal cut-point value in roc analysis: An alternative approach. *Comput Math Methods Med*.

Ward, A. B., Sali, A., and Wilson, I. A. (2013). Integrative Structural Biology. *Science*, 339:913–915.

STAR ⋆ METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
NRGXL server	This paper	https://nrgxl.pasteur.fr
NRGXL software and manual	This paper	https://gitlab.pasteur.fr/bougui/NRGXL
CNS	(Brunger et al., 1998; Brunger, 2007)	http://cns-online.org/v1.3/
Other		
XLdb	(Kahraman et al., 2013)	https://doi.org/10.1371/journal.pone.0073411.s008

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for data should be directed to and will be fulfilled by the lead contact Guillaume Bouvier (guillaume.bouvier@pasteur.fr).

METHOD DETAILS

Sampling approach

NRGXL samples cross-linker conformations between two reactive amino-acid residues, usually lysines. For each cross-link, the system sampled is composed of the side chains of the two residues and the cross-linker molecule bridging them. The cross-linker and the lysine side chains are incrementally built step by step from the C_β of the cross-linked lysines.

Before sampling, the initial protein structure is prepared as follows: i) the protein backbone is fixed; ii) the side chains of the two residues are reduced to their C_β atoms; iii) the C_β of the two residues are selected as start and endpoints of the incremental building; iv) the number of incremental building steps is established by the length of the cross-linker and the length of the two side chains; v) a grid-based density map is created surrounding the entire macro-molecular complex, the grid points being used to place atoms of the sampled conformations vi) for each pair of cross-linked residues, a smaller grid box is selected surrounding the two residues, and the transition matrix and its powers are calculated (see Density map and transition matrix).

After this preparation stage, conformations are generated one by one. The incremental building of the cross-linker and the lysine side chains are driven by a constrained Markov Chain algorithm (see Constrained Markov Chain algorithm), which first construct the side chain of the start residue, then the cross-linker molecule and finally the side chain of the end residue. Concretely, the constrained Markov chain algorithm samples a path of fixed length between the fixed start and endpoints. This path is generated step by step, moving at each step from one grid point to another. The sampling movement is directed by the transition probabilities, which provide the probabilities of moving from one grid point to another, given the current position and the initial conditions. The transition probability is higher when the density at a grid point is lower in such a way that paths avoiding the protein matrix are preferentially sampled.

Once a conformation of the cross-linker and the attached side chains is generated, we use CNS (Brunger et al., 1998; Brunger, 2007) to energy minimize the conformation and obtain an overall energy value for its interaction with the protein and its internal conformation (for more details see CNS minimization).

This process is repeated several times, generating an ensemble of different cross-linker and side chain conformations, together with their energies. Having multiple conformations for a given pair of residues serves to choose an optimal conformation, i.e., the one of lowest energy, and also serves as a measure of the diversity of the possible conformations.

Density map and transition matrix

The density map is computed from a given PDB structure by making use of the Python library TEMPy (Farabella et al., 2015). It is generated as a mesh grid with the step size of 1.1 Å. This value results in an average distance over all possible directions of the grid of about the length of a C-C bond (1.54 Å).

The constrained Markov Chain algorithm is characterized by its transition matrix, which works as a measure for the transition probabilities (see Constrained Markov Chain algorithm). This transition matrix is square $n \times n$ where n is the number of cells in the density grid. An element a_{ij} of the matrix represents the probability measure (after normalization) between the cell i and the cell j and is given by:

$$a_{ij} = P_{ij} = \begin{cases} e^{-\beta(d_j - d_i)}, & \text{if } i \text{ and } j \text{ are neighbours in space} \\ 0, & \text{if } i \text{ and } j \text{ are not neighbours in space} \end{cases} \quad (1)$$

where d_j and d_i are the densities of the cells j, and i, respectively, and β is a parameter which works as a modulator of the exponential intensity (in our study it was set to 50 after a short process of optimization).

We chose this probability in order to give a sense of directionality to the sampling process. If $d_j \gg d_i \Rightarrow d_j - d_i \gg 0 \Rightarrow P_{ij} \to 0$ meaning that the probability to move from a cell with low density to one with high density is very low (as it should be).

On the other hand, if $d_j \ll d_i \Rightarrow d_j - d_i \ll 0 \Rightarrow P_{ij} \to \infty$, this implies that the probability of moving from a cell with higher density to another with lower density is quite high. This directionality prioritizes the formation of side-chain and cross-linker conformations in areas with low density.

We work with small grids around each cross-linked residue pair to drastically reduce memory requirements, in particular in large macro-molecular complexes. The small grid has the size necessary to encompass all possible cross-linker conformations. The transition matrix is then computed in this small box rather than on the whole density grid.

Constrained Markov Chain algorithm

We use a Constrained Markov Chain to sample conformations of the cross-linker, and the two side chains Constrained Markov Chain. This is a stochastic process (collection of random variables X_n) that satisfies the Markov property (the conditional probability distribution of future states of the process depends only on the present state). A constrained Markov Chain can be defined as a Markov Chain in which for every step the transition probabilities have a new constant condition. For a given probability space, with P_{ij} a probability measure and $\{X_i\}_{i\in\mathbb{N}}$ a set of random variables, the transition probability between states n and $n+1$ is:

$$P(X_{n+1} = b | X_n = a, X_N = c) = \frac{P_{bc}^{N-(n+1)} \cdot P_{ab}}{P_{ac}^{N-n}} \tag{2}$$

where $X_N = c$ is the constant condition for each step.

When sampling cross-linker conformations between a pair of reactive residues, the constant condition $X_N = c$ of the algorithm serves to enforce the cross-linked structure to finish at a given position c (the fixed end point) and at a given state number N (fixed length of the cross-link). Moreover, the probability measure P_{ij} is given by the transition matrix (see Density map and transition matrix). The detailed proof of Eq.2 is given in the Constrained Markov Chain algorithm proof subsection of the Supplemental Information.

The fact that the probability measure P_{ij} depends on the density at the grid point leads in some cases to the result that fewer conformations than desired, or no conformations at all, are found. The transition probabilities between one state and its spatial neighbour states are calculated according to equations Eq.2 and Eq.1 (the states that are not neighbours in space have transition probability 0). The states of the sampling process are selected step by step, according to these probabilities. If all neighbor transition probabilities are lower than the parameter ε that Python uses to determine if a particular value is 0, the transition probability vanishes. In this case, the next stage in the sampling process cannot be selected, and the process ends without being able to reach the fixed endpoint (the position of the C_β of the end residue). These incomplete cross-linker conformations are discarded from the final ensemble, generating ensembles with less than 20 conformations.

Constrained Markov Chain algorithm proof

Definition The m-step transition probability is the probability to move from state i to state j in m steps:

$$p_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

Definition The *single*-step transition probability is the probability to move from state i to state j in one step:

$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

Definition The m-step transition matrix $P^{(m)}$, is the matrix whose $\{a_{ij}\}$ elements are the m-step transition probabilities.
Definition The *single*-step transition matrix P, is the matrix whose $\{a_{ij}\}$ elements are the *single*-step transition probabilities.

Lemma 1 *The m-step transition matrix, is equal to the single-step transition matrix multiplied by itself m times.*

$$P^{(m)} = P^m$$

Proof: Moving from state i to state j in m steps is the same than moving first from state i to state r in m-k steps and then from state r to state j in k steps:

$$p_{ij}^{(m)} = \sum_r p_{ir}^{(m-k)} p_{rj}^{(k)},$$

which implies that

$$P^{(m)} = P^{(m-k)} \cdot P^{(k)}.$$

Setting $k = m-1$ we have that

$$P^{(m)} = P \cdot P^{(m-1)}, \tag{3}$$

from which we can deduce that

$$P^{(m-1)} = P \cdot P^{(m-2)}. \tag{4}$$

Substituting Eq.4 into Eq.3 we have that

$$P^{(m)} = P \cdot P \cdot P^{(m-2)},$$

and iterating

$$P^{(m)} = P \cdot P \cdots P = P^m$$

Theorem 1 *The single-step transition probability of a Markov Chain between the state i and the state j knowing that in N steps we must be at state i_N is:*

$$P(X_{n+1} = j | X_n = i, X_N = i_N) = \frac{P_{ji_N}^{N-(n+1)} \cdot P_{ij}}{P_{ii_N}^{N-n}}$$

Proof:

$$P(X_{n+1} = j | X_n = i, X_N = i_N) = \frac{P(X_{n+1} = j, X_n = i, X_N = i_N)}{P(X_n = i, X_N = i_N)} =$$

$$\frac{P(X_N = i_N | X_{n+1} = j, X_n = i) P(X_{n+1} = j, X_n = i)}{P(X_n = i, X_N = i_N)} =$$

$$\frac{P(X_N = i_N | X_{n+1} = j) P(X_{n+1} = j | X_n = i) P(X_n = i)}{P(X_N = i_N | X_n = i) P(X_n = i)} =$$

$$\frac{P(X_N = i_N | X_{n+1} = j) P_{ij}}{P(X_N = i_N | X_n = i)} =$$

$$\frac{P_{ji_N}^{N-(n+1)} P_{ij}}{P_{ii_N}^{N-m}}$$

where in the third equality we used the Markov property ($P(X_N = i_N | X_{n+1} = j, X_n = i) = P(X_N = i_N | X_{n+1} = j)$) and in the last we used the Lemma 1

CNS minimization

To rapidly minimize the energy of the generated cross-linker conformations, we use the program CNS (Brunger et al., 1998; Brunger, 2007). To generate the topology file for the used cross-linker, we used the PRODRG server (Schüttelkopf and van Aalten, 2004). The topology defines the different bonded and non-bonded interactions between specific atoms through a force field (CHARMM (MacKerell et al., 1998)), which is used to compute potential energy. The total energy includes internal parameters (bond length, bond angle, improper and dihedral angles) and non-bonded interactions with full van der Waals and electrostatic potentials with a 13 Å non-bonded cutoff. Cross-linker conformations were minimized with 2,000 steps of conjugate gradient minimization. During the minimization, atoms of the protein complex were kept fixed, and only the atoms of the cross-linker and cross-linked side chains could move freely.

Anisotropic Network Model

To assess the influence of the global flexibility of the entire system, we used normal mode analysis with an anisotropic network model (Atilgan et al., 2001) (ANM) with ProDy (Bakan et al., 2011) (open-source Python package for protein structural dynamics). ANM nodes were the C_α atoms. For each entry in the XLdb database, ten decoy conformations were generated along the first non-trivial mode, with a maximum amplitude of oscillation of 2 Å.

QUANTIFICATION AND STATISTICAL ANALYSIS

Binary Classification study

The binary classification scheme serves to predict if two residues can be cross-linked based on several features. It served us to rigorously compare the quality of predictions based on the different features. For this, we analyzed all lysine pairs from a set of 53 different PDB structures for which experimental XL-MS data were available in the XLdb database (Kahraman et al., 2013) (see XLdb) that were evaluated as compatible by a simple distance criterion. This database contains cross-linking data using BS3/DSS cross-linkers. Size, hydrophobicity, and charge of the cross-linker previous to the cross-linking event were not considered in our analysis. Therefore, due to the identical conformation of DSS and BS3 after being cross-linked, we equally handle both types of cross-links. In summary, we analyzed lysine pairs with C_β-C_β distance shorter than 28.42 Å, which is the maximum length that an XL conformation formed by two lysine side chains (13.44 Å each), and a BS3/DSS cross-linker (14.98 Å), can adopt.

Database entries with high inconsistency between experimental data (list of cross-links from XLdb) and structural data (PDB structures) were discarded. More precisely, for all systems, a ROC plot per feature was computed. Then, systems with an AUC below 0.75 for all features were removed. In other words, we ended up discarding database entries for which structure and experimental data were not in agreement for any feature.

During the binary classification analysis, the same number of compatible lysine pairs with an experimentally detected XL (EyXL) and compatible lysine pairs without an experimentally detected XL (EnXL) were selected to ensure balanced population sets. For both the *static* and *flexible models*, we randomly selected compatible lysine pairs to arrive at equal numbers of EnXL and EyXL. For all compatible lysine pairs in the selected population (EyXL and EnXL), features such as SASD, EucDist, or NRGXL energy were computed.

Each computed feature value was tested as a classifier. For a given threshold, compatible lysine pairs with a value smaller (or higher depending on the physical meaning of the feature) than this threshold were classified as cross-linkable (CyXL), and compatible lysine pairs with a value higher (or smaller, respectively) than the threshold were classified as not cross-linkable (CnXL). EyXL classified as CyXL are considered as true positives (TP), and EyXL classified as CnXL we labeled as false negative (FN). Accordingly, False positives (FP) and true negatives (TN) correspond to EnXL classified as CyXL or CnXL, respectively. To check the accuracy and precision of our binary classification process for a specific feature threshold as a classifier, we used the following equations:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

The computed accuracy reports the proportion of well-classified compatible lysine pairs while the precision reports the proportion of positive classified pairs (CyXL) that are genuine positives (CyXL and EyXL simultaneously).

For each feature, and multiple thresholds, we computed the value of accuracy, the value of precision, the false-positive rate ($FPR = \frac{FP}{FP+TN}$) and the true-positive rate ($TPR = \frac{TP}{TP+FN}$). Accuracies were first computed during a training step in order to define an optimal threshold (as the one with the highest accuracy). Then these optimal thresholds were employed during a validation step to calculate a new accuracy and a precision value. In contrast, the FPR and TPR ratios were measured at once (without partitioning the data into training and validation) to generate a Receiver Operating Characteristic (ROC) curve. Accuracy, precision and ROC curves served to compare the quality of different features as a classifier.

Binary Classification study: treatment of random errors

To obtain balanced population sets of EyXL and EnXL, we randomly picked a group of EnXL to match the number of EyXL. This stochastic selection could induce random errors. For this reason, for each feature threshold, we generated 100 random balanced populations and computed their respective accuracies, precisions, FPR, and TPR. Then, for the particular threshold, the mean accuracy, precision, FPR, and TPR over these 100 populations were selected as representative. In addition, to ensure that the obtained mean values of accuracy per feature threshold in the training set were not affected by the random selection of balanced populations, we computed a p-value per each one of them. For this purpose, random average accuracies were calculated in parallel to the usual average accuracies and then compared. These average random accuracies were calculated by arbitrarily associating a value of cross-linkability (EyXL or EnXL) to all the cross-links involved in the binary classification study. In this way, we perform parallel binary classification studies for a random classifier from which we obtain random average accuracies. Thereby, we end up with a random and a regular average accuracy per feature threshold which can be compared to obtain a p-value used as a criterion to define two twilight zones (for low and high thresholds respectively) where the accuracy distribution of the random classifier overlaps with our classifier.

DATA AND CODE AVAILABILITY

XLdb

XLdb (Kahraman et al., 2013) is a cross-link database, containing 506 intra-protein and 62 inter-protein cross-links from 14 publications. For all the entries in the database, either DSS or BS3 cross-linkers were used, which implies that all the database cross-links are between lysine residues. All these cross-links can be mapped to 56 different PDB structures. All the structures and the list of residue pairs cross-linked used are publicly available in the NRGXL repository (see below).

NRGXL implementation

NRGXL is implemented in Python and is freely available as a python package. The most recent version can be found here: https://gitlab.pasteur.fr/bougui/NRGXL. NRGXL is also accessible as a web server at: https://nrgxl.pasteur.fr.

Figures & Tables titles/legends

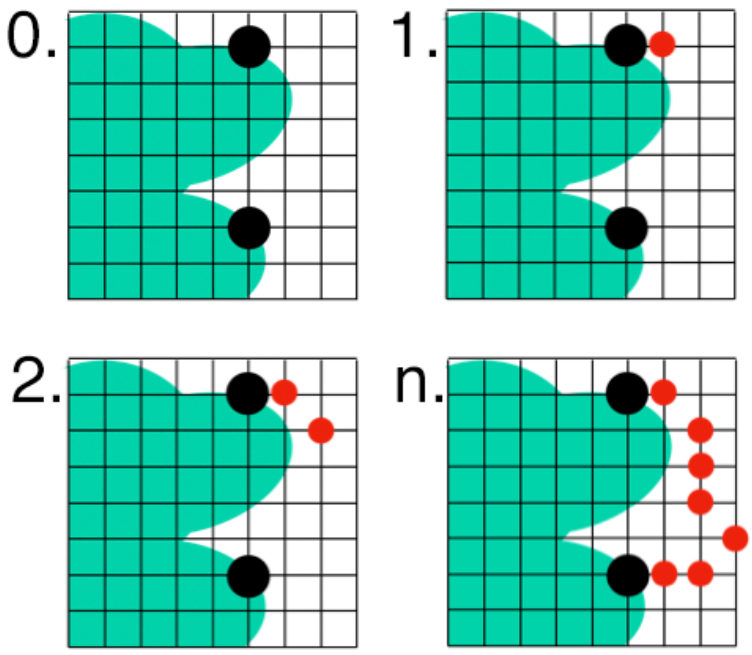


Figure 1. Sampling approach
Sampling algorithm 2D (from 0 to *nth* step) sketch used to generate cross-link conformations in 3D grids. Black dots represent the grid cells where the C_β of the cross-linked residues are located and hence the Constrained Markov Chain initial conditions. Each red dot represents one sampling step.

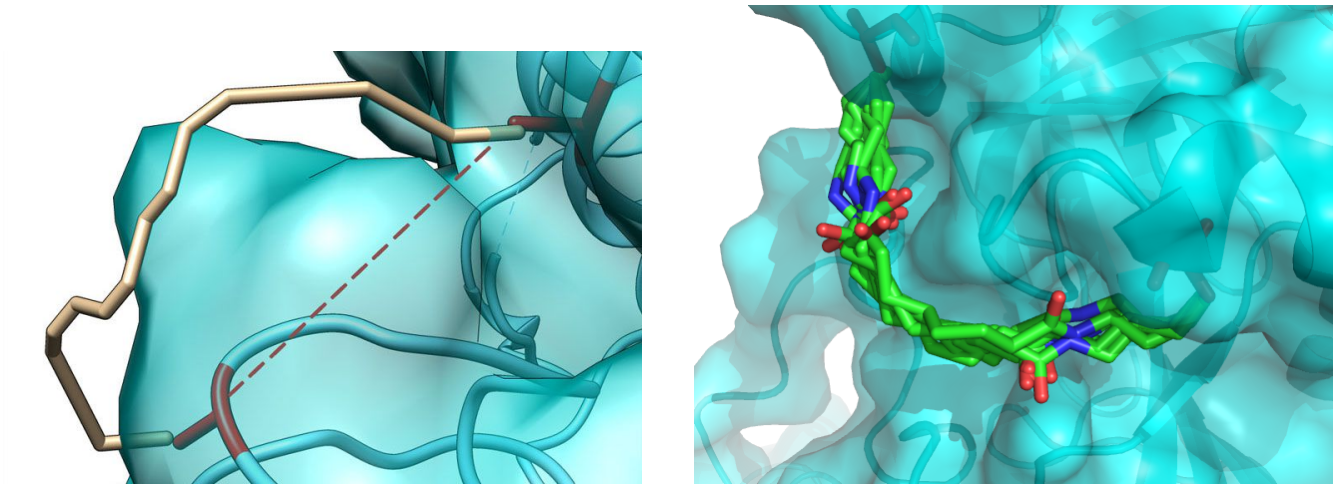


Figure 2. Different approaches to computationally characterize cross-linker structures
(a) Euclidean distance and shortest solvent-accessible surface distance between two lysines in RNA Pol II, both computed between the C_β of the two residues. (b) An ensemble of sampled cross-link conformations between two lysines in RNA pol II, starting and ending in C_β atoms. This ensemble was computed with our approach, which associates conformational energy to each of the cross-link conformations.

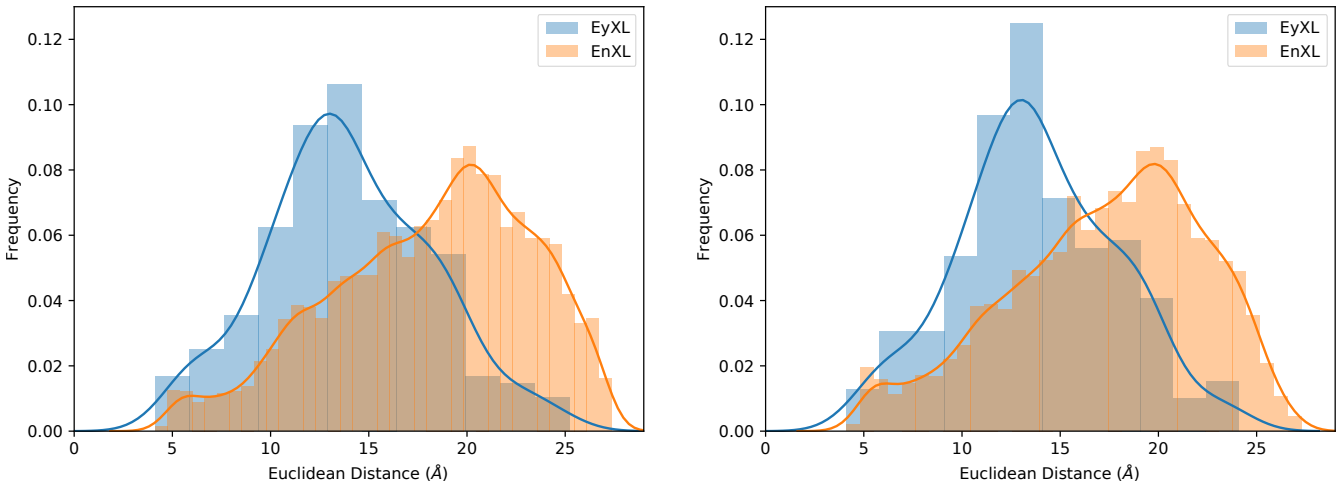


Figure 3. Euclidean distance distributions
(a), (b) Comparison of EucDist values between EyXL and EnXL for the *static* (a) and the *flexible model* (b)

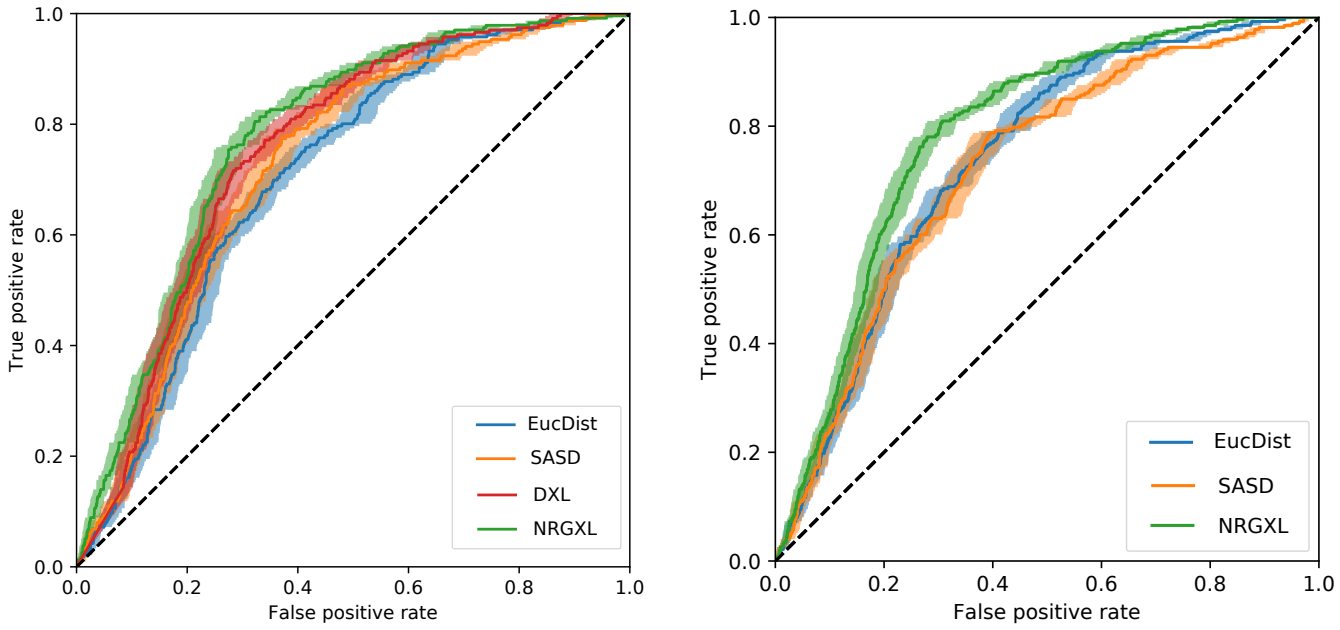


Figure 4. ROC plots for different statistics
(a), (b) To generate each one of these ROC plots, different values of the features were selected as thresholds during a Binary Classification study. Then, for each one of them, their mean false positive rate and their mean true positive rate were computed over 100 randomly selected EyXL/EnXL balanced populations. The light-colored error bars surrounding the curves represent their $\pm$ standard deviation. In figure (a), the ROC plots represent features of the *static model*. Meanwhile, the ROC plots of figure (b) represent features of the *flexible model*.

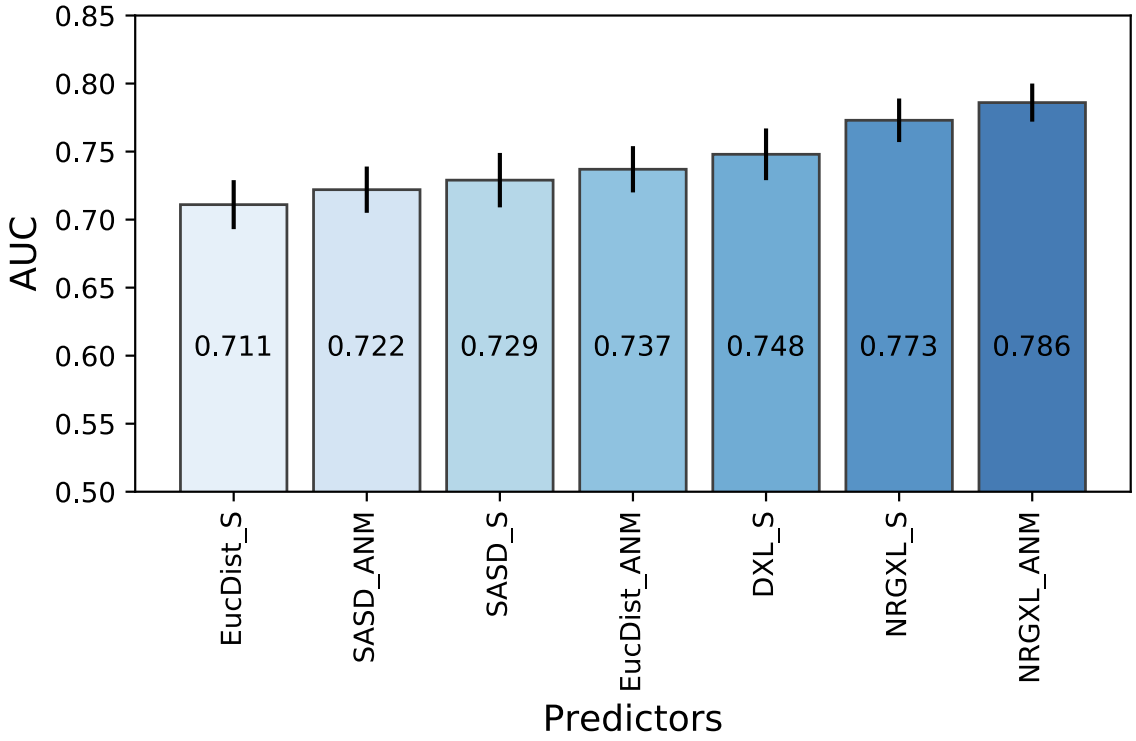


Figure 5. AUC from different statistics and different models
These values where obtained from the ROC plots of the Figure 4. S and ANM are related with the *static model* and the *flexible model* respectively.

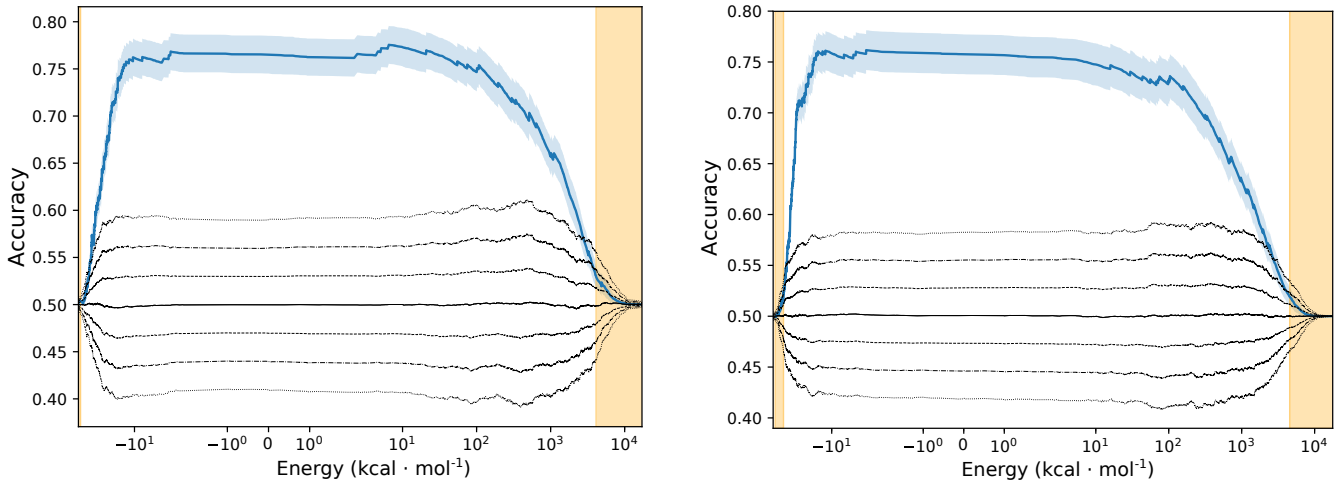


Figure 6. Mean accuracy for different energy thresholds of both the *static* and the *flexible model* on a training subset. (a), (b) The blue plot represents the mean accuracy per energy threshold obtained over 100 randomly selected EyXL/EnXL balanced populations. The light blue error bars surrounding the mean accuracy plot represent the $\pm$ standard deviation. The black plot, almost perfectly aligned with *accuracy* $= 0.5$, represents the generated random mean accuracy per energy threshold computed in order to calculate mean accuracy p-values (see Binary Classification study: treatment of random errors). The different dotted lines represent one, two, and three times the random standard deviations. Finally, orange-colored areas show energy thresholds for which the corresponding accuracy has a p-value below 0.05, and which are therefore not significant. (a) *static model*; (b) *flexible model*.

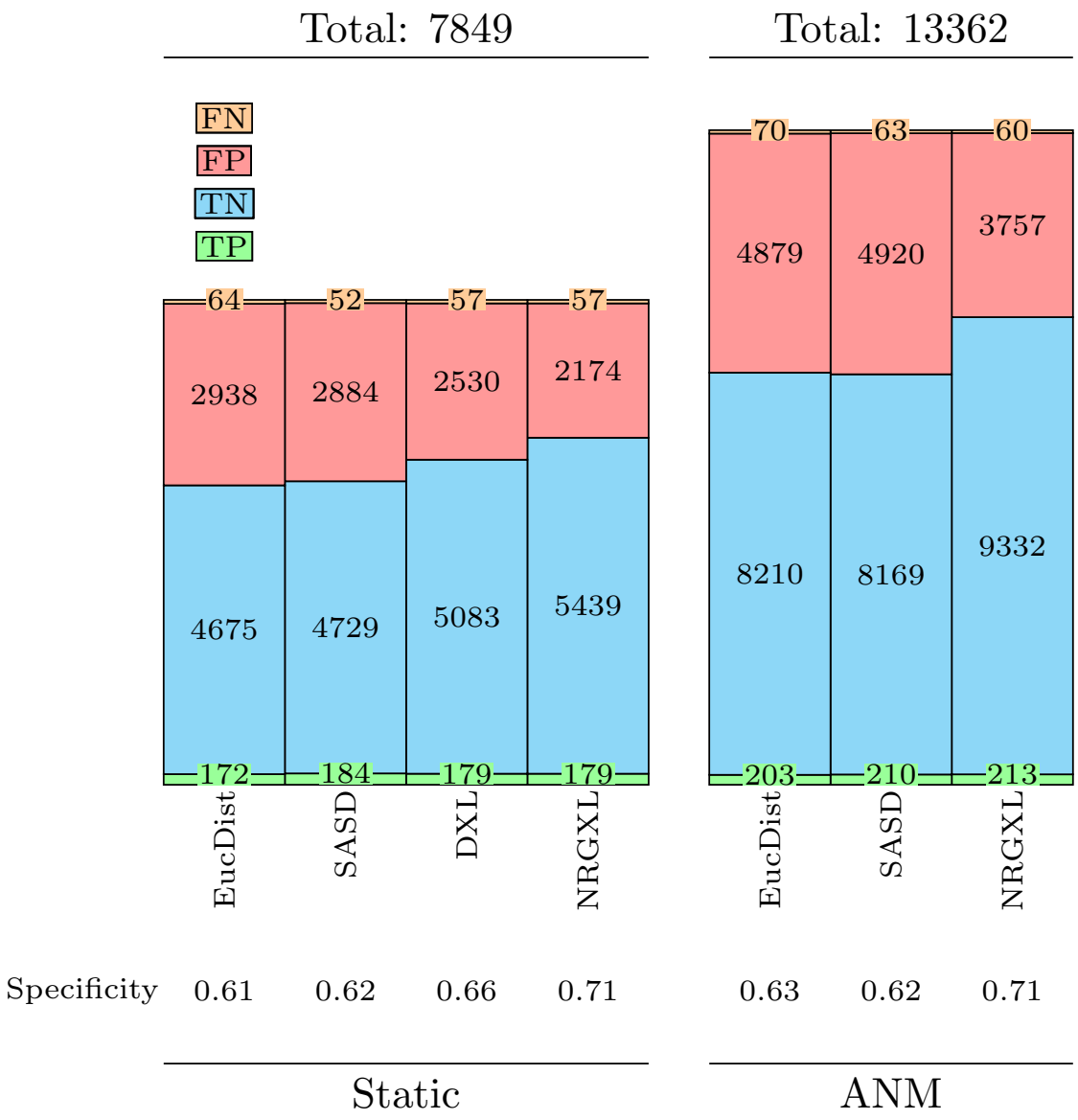


Figure 7. Plain outcomes for Binary Classification studies.
Outcomes of seven Binary Classification studies for the different features available in both models and using all compatible lysine pairs (without balanced populations). The threshold for each of these studies was set to the range of the optimal thresholds. The resulting specificity per each study is shown below. Static and ANM refer to the *static model* and the *flexible model* described previously. The values indicated on the bar plot give the count for the four corresponding classes (FN, FP, TN, TP).

	Training		Validation	
	Accuracy	Optimal Threshold	Accuracy	Precision
EucDist_S	0.68 ± 0.02	16.03 ± 0.95 Å	0.66 ± 0.02	0.65 ± 0.03
SASD_S	0.71 ± 0.02	30.59 ± 1.33 Å	0.69 ± 0.03	0.67 ± 0.03
DXL_S	0.72 ± 0.02	9.08 ± 0.80 Å	0.71 ± 0.02	0.70 ± 0.03
NRGXL_S	0.75 ± 0.02	1.35 ± 12.19 kcal $\cdot$ mol^{-1}	0.74 ± 0.02	0.72 ± 0.03
EucDist_ANM	0.70 ± 0.02	16.80 ± 1.35 Å	0.67 ± 0.02	0.66 ± 0.02
SASD_ANM	0.70 ± 0.02	30.28 ± 0.35 Å	0.69 ± 0.03	0.67 ± 0.03
NRGXL_ANM	0.76 ± 0.02	-8.83 ± 5.91 kcal $\cdot$ mol^{-1}	0.75 ± 0.02	0.73 ± 0.03

Table 1. Training highest mean accuracy, optimal threshold, and its corresponding validation average accuracy and average precision per feature and model.
S and ANM stand for *static* and *flexible model*

	Sampling time	Preparation Time	Overall Time
1wcm (4521 residues)	0.479 ± 0.050s	0.330 ± 0.207s	0.809 ± 0.146s
4f5s (1166 residues)	0.479 ± 0.048s	0.133 ± 0.037s	0.612 ± 0.056s
4fgf (146 residues)	0.502 ± 0.060s	0.022 ± 0.001s	0.524 ± 0.060s

Table 2. Efficacy study results
Efficiency test outcomes from three different sized complexes presented as the mean generation time per cross-link conformation $\pm$ standard deviation computed through all the compatible lysine pairs per system.

Supplemental Information

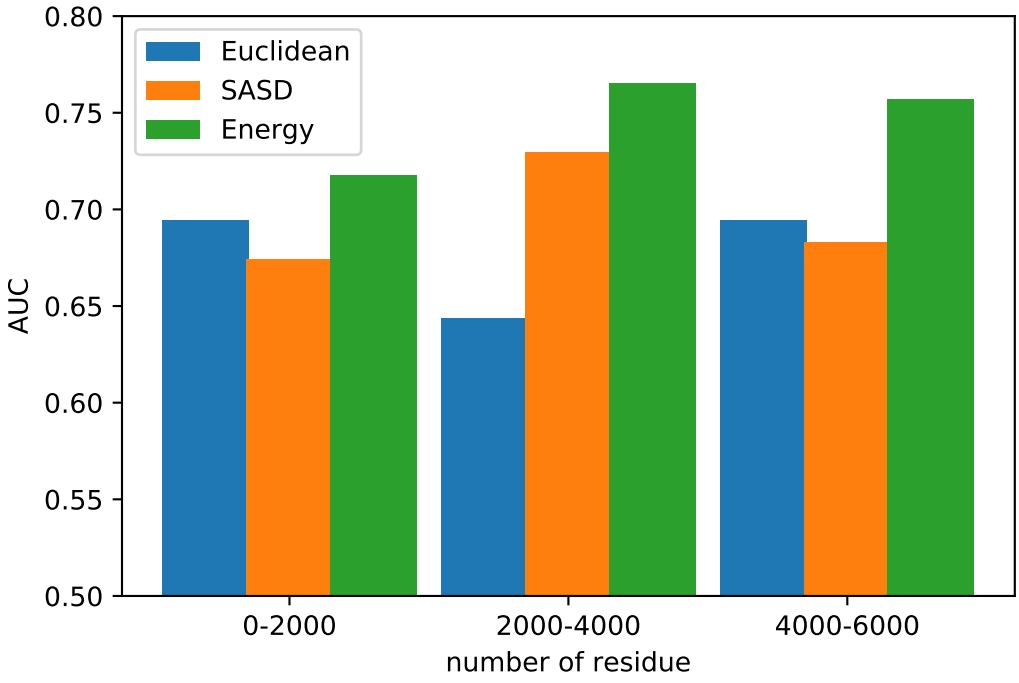


Figure S1. Effect of the protein size on different classifier performance. Related to Figure 5
Compatible lysine pairs from the static model were splitted into 3 groups based on the number of residues of the protein were they belong (0-2000, 2000-4000, 4000-6000 residues). Then, we computed an AUC per classifier (Eulidean, SASD or energy) for each defined group.

	Static				ANM		
	EucDist_S	SASD_S	DynamXL_S	NRGXL_S	EucDist_ANM	SASD_ANM	NRGXL_ANM
Accuracy Train.	0.68 ± 0.02	0.71 ± 0.02	0.72 ± 0.02	0.75 ± 0.02	0.70 ± 0.02	0.70 ± 0.02	0.76 ± 0.02
Accuracy OT. Train.	16.03 ± 0.95 Å	30.59 ± 1.33 Å	9.08 ± 0.80 Å	1.35 ± 12.19 kcal $\cdot$ mol^{-1}	16.80 ± 1.35 Å	30.28 ± 0.35 Å	-8.83 ± 5.91 kcal $\cdot$ mol^{-1}
Accuracy Val.	0.66 ± 0.02	0.69 ± 0.03	0.71 ± 0.02	0.74 ± 0.02	0.67 ± 0.02	0.69 ± 0.03	0.75 ± 0.02
Precision Val.	0.65 ± 0.03	0.67 ± 0.03	0.70 ± 0.03	0.72 ± 0.03	0.66 ± 0.02	0.67 ± 0.03	0.73 ± 0.03
Sensitivity Val.	0.69 ± 0.06	0.77 ± 0.06	0.73 ± 0.04	0.78 ± 0.04	0.72 ± 0.09	0.76 ± 0.04	0.78 ± 0.04
Specificity Val.	0.63 ± 0.08	0.60 ± 0.06	0.69 ± 0.04	0.70 ± 0.05	0.62 ± 0.08	0.63 ± 0.04	0.72 ± 0.04
NPV Val.	0.68 ± 0.03	0.74 ± 0.04	0.72 ± 0.03	0.76 ± 0.03	0.70 ± 0.05	0.72 ± 0.03	0.76 ± 0.03
(0,1) Distance Train.	0.46 ± 0.02	0.42 ± 0.01	0.39 ± 0.02	0.36 ± 0.02	0.43 ± 0.02	0.43 ± 0.01	0.35 ± 0.01
(0,1) OT. Train.	15.58 ± 0.49 Å	29.34 ± 0.64 Å	8.67 ± 0.54 Å	-6.78 ± 8.03 kcal $\cdot$ mol^{-1}	15.94 ± 0.41 Å	29.74 ± 0.61 Å	-12.31 ± 3.87 kcal $\cdot$ mol^{-1}
Accuracy Val.	0.66 ± 0.02	0.69 ± 0.03	0.70 ± 0.02	0.73 ± 0.02	0.68 ± 0.02	0.69 ± 0.02	0.74 ± 0.02
Precision Val.	0.66 ± 0.03	0.68 ± 0.02	0.70 ± 0.03	0.73 ± 0.03	0.68 ± 0.03	0.67 ± 0.02	0.73 ± 0.03
Sensitivity Val.	0.67 ± 0.05	0.73 ± 0.06	0.72 ± 0.04	0.76 ± 0.04	0.69 ± 0.04	0.72 ± 0.05	0.76 ± 0.03
Specificity Val.	0.66 ± 0.05	0.66 ± 0.04	0.69 ± 0.05	0.71 ± 0.05	0.68 ± 0.05	0.65 ± 0.04	0.72 ± 0.04
NPV Val.	0.67 ± 0.03	0.70 ± 0.04	0.71 ± 0.03	0.74 ± 0.03	0.68 ± 0.03	0.69 ± 0.04	0.75 ± 0.02

Table S1. Training highest mean accuracy, optimal threshold, and corresponding validation statistical metrics per feature and model obtained through two different threshold selection approaches. Related to Table 1.
Extension of Table 1 with the closest-to-(0,1) threshold selection alternative (Unal, 2017) and with the inclusion of three new statistical measures (sensitivity, specificity and negative prediced value) per optimal cutting point method. Closest-to-(0,1) approach utilise a ROC curve to calculate a feature threshold. This is defined as the curve point minimizing its straight-line distance with the (0,1) corner. S and ANM stand for *static* and *ANM* model respectively, OT for Optimal Threshold, NPV for negative predicted value ,Train. for training and Val. for validation.