Computational studies of molecular pre-organization through macrocyclization: Conformational distribution analysis of closely related non-macrocyclic and macrocyclic analogs

Gustav Olanders, Peter Brandt, Christian Sköld, Anders Karlén

Department of Medicinal Chemistry, Uppsala University, BMC, Box 574, SE-751 23 Uppsala, Sweden
Medicinal Chemistry, Research and Early Development Cardiovascular, Renal and Metabolism, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

ARTICLE INFO
Keywords:
Macrocycle
Macrocyclization
Linear
Pre-organization
Rigidification
Structure-based design
Conformational properties

ABSTRACT

Macrocycles form an important compound class in medicinal chemistry due to their interesting structural and biological properties. To help design macrocycles, it is important to understand how the conformational preferences are affected upon macrocyclization of a lead compound. To address this, we collected a unique data set of protein–ligand complexes containing “non-macrocyclic” (“linear”) ligands matched with macrocyclic analogs binding to the same protein in a similar pose. Out of the 39 co-crystallized ligands considered, 10 were linear and 29 were macrocyclic. To enable a more general analysis, 128 additional ligands from the publications associated with these protein data bank entries were added to the data set. Using in total 167 collected ligands, we investigated if the conformers in the macrocyclic conformational ensembles were more similar to the bioactive conformation in comparison to the conformers of their linear counterparts. Unexpectedly, in most cases the macrocycle conformational ensemble distributions were not very different from those of the linear compounds. Thus, care should be taken when designing macrocycles with the aim to focus their conformational preference towards the bioactive conformation. We also set out to investigate potential conformational flexibility differences between the two compound classes, computational energy window settings and evaluate a literature metric for approximating the conformational focusing on the bioactive conformation.

1. Introduction

Molecular rigidification restricts the number of conformers a ligand can adopt and is an important tool in medicinal chemistry when developing more potent and selective analogs. The rationale behind rigidifications is often to improve affinity and selectivity and to lock the ligand in its bioactive conformation to facilitate target binding while avoiding conformations that allow for interactions with other targets. Accordingly, rigidification through macrocyclization as a tool to improve binding, bioavailability, toxicity, and other properties, is continuing to gain attention. 1–8 Macrocycles are often said to be conformationally pre-organized, and are defined as compounds containing a ring of at least 8–12 ring atoms. 2,4,9,10 The term pre-organized is commonly used to describe a compound where the conformational ensemble favors conformers which put important pharmacophore groups in positions beneficial for binding, i.e. the compound is rigidified to mimic its bioactive conformation. Today there are several macrocyclic compounds marketed for clinical use where most are natural products or derivatives thereof. 2,4 Given the synthetic challenge associated with macrocycles, specialized protocols for docking and conformational sampling are emerging to aid the design process. 11–18 Additionally, automatic linker design tools and linker evaluation metrics are being used. 17 Recently, a Boltzmann averaged root mean squared deviation (RMSD) metric has been proposed for linker evaluations. 19 Commonly, the premier focus of the computational linker design evaluation is to approximate which of the considered linkers will most effectively focus the conformational ensemble on the bioactive conformation. Since macrocyclization can be used to steer the conformational ensemble, understanding how the conformational ensemble is affected is important. Thus, to rationally design new synthetic macrocycles more studies on how macrocyclization affects the conformational distribution is needed. As mentioned earlier, conformational analysis can be used as a tool to evaluate if different linkers enable the adoption of the bioactive conformation. 15,19,20 In the current work, we set out to investigate if

* Corresponding author.
E-mail address: anders.karlen@ilk.uu.se (A. Karlén).

https://doi.org/10.1016/j.bmc.2021.116399
Received 5 May 2021; Received in revised form 24 August 2021; Accepted 1 September 2021
Available online 11 September 2021
0968-0896/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
synthetic macrocyclic analogs populate conformations closer to the X-ray-bound conformation to a higher degree as compared to their non-macrocyclic counterparts (henceforth called “linear”). We also set out to investigate how the macrocyclization affects conformational flexibility and how the purposed linker design metric performs on our collected data set.

To facilitate this study, a data set was needed. In an early medicinal chemistry project, typically only an X-ray structure containing a linear ligand would be present and the proposed macrocycles would be assumed to bind in a similar pose. In the literature, several macrocycle containing data sets can be found. However, to evaluate if the macrocycles favor the bioactive conformation, X-ray conformers of both linear ligands and at least one macrocyclic analog binding to the same protein target are needed to validate that they bind in a similar pose. To the best of our knowledge, few such data sets are available in the literature today. Thus, we compiled a data set of X-ray structures from the protein data bank (PDB) comprised of linear compounds and macrocyclic analogs binding to the same protein target in a similar binding mode, making similar protein-ligand interactions. To expand the data set, close analogs from the original publications were also included.

2. Methods

All calculations were performed using Schrödinger Small-Molecule Drug Discovery Suite, Schrödinger (Schrödinger Research 2019–3) using the OPLS3 force field and a GB/SA continuum solvation model for water. All plots were made in Python, whereas all molecular modeling figures were made in Maestro.

2.1. Data set selection

First we added targets based on our previous publications and prior knowledge of which proteins have been targeted with macrocycles. Other targets were found by PDB title searches. However, most ligands were found by downloading all ligands from PDB entries with a 2.5 Å resolution or better (downloaded multiple times from 2017 to 2020). The number of ring atoms was thereafter calculated using ChemAxon and all ligands containing a ring with 10 ring atoms or more were visually inspected. Out of this selection, only synthetic and pharmacologically relevant macrocyclic ligands were kept and compounds such as crown ethers, chelators, sugars, and natural products such as erythromycin and tacrolimus, were removed. Linear analogs binding to the same protein were thereafter identified by using the protein sequence similarity function in PDB and substructure search using DataWarrior. Binding data were collected from the original publications, when binding data was not available in the original papers, PubChem was used. See Supporting Information for detailed data origin.

2.2. Data set preparation

The PDB protein–ligand complexes were prepared using the Protein Preparation Wizard in Maestro with default options. This included the addition of missing side-chains, management of cases where amino acid residues had alternate positions (first listed or highest average occupancy were selected), generation of ligand tautomers, and ionization states, as well as protein protonation states and, hydrogen bond network optimization. The energy minimization step in the Protein Preparation Wizard was not performed. Before including a structure into the data set, the fit to the electron density was evaluated qualitatively for all protein–ligand complexes using maps from the Uppsala electron density server or the PDB.

2.3. Generating the conformational ensembles and identifying the lowest energy minimum

The conformational ensemble was generated by pooling the results of two consecutive conformational searches. The first conformational search was performed using 10,000 steps MCMM (Monte Carlo Multiple Minimum) with previously published enhanced macrocycle sampling settings starting from the conformation generated from the SMILES string of the compound. The second conformational search consisted of 20,000 steps MD/LLMOD (macrocycle sampling in MacroModel) starting from the lowest energy conformer generated from the MCMM run. The conformations generated by MCMM and MD/LLMOD were pooled together and submitted for up to 50,000 steps of energy minimization using the truncated Newton conjugate gradient (TNCG) method with a gradient convergence criterion of 0.01 kJ Å⁻¹ mol⁻¹ with a 15 kcal mol⁻¹ energy window and duplicate conformers removed. As in the MCMM conformational search, conformers were considered unique if one heavy atom deviated more than 0.5 Å from the other conformers after superposition. The lowest energy minimum identified for a compound is defined as the lowest energy conformation found after energy minimizing the two pooled conformational searches, hereafter called “global energy minima”.

2.4. Minimization of the X-ray ligand structures

The protein-bound ligand conformations were extracted from the complexes generated by the Protein Preparation Wizard (after the Hydrogen bond network optimization step) and minimized with MacroModel in three steps. Firstly, all torsions were constrained with a force constant of 1,000 kJ mol⁻¹ deg⁻². Secondly, all ligands were energy minimized in a constrained manner using a force constant of 500 kJ mol⁻¹ Å⁻². Finally, the ligands were minimized without any constraints. The TNCG energy minimization method was used with a maximum of 50,000 iteration steps and a gradient convergence criterion of 0.01 kJ Å⁻¹ mol⁻¹ in all steps.

2.5. Root mean square deviation calculations

The Root Mean Square Deviation (RMSD) calculations were performed using the superposition tool in Maestro. The RMSD values were calculated using the heavy atoms of the structural elements shared between the linear compounds and the macrocyclic analogs as annotated in Table 1. The corresponding linear X-ray ligand was used as the template for superposition. In cases of element mismatch in part of the structures, e.g. phenyl versus pyridine, between the linear ligand and its macrocyclic analogs, these atomic positions were still included in the RMSD calculations and are therefore marked in blue in Table 1. The RMSD values were calculated using the heavy atoms of the structural elements shared between the linear compounds and the macrocyclic analogs as annotated in Table 1. The corresponding linear X-ray ligand was used as the template for superposition. In cases of element mismatch in part of the structures, e.g. phenyl versus pyridine, between the linear ligand and its macrocyclic analogs, these atomic positions were still included in the RMSD calculations and are therefore marked in blue in Table 1. The RMSD values were calculated using the heavy atoms of the structural elements shared between the linear compounds and the macrocyclic analogs as annotated in Table 1. The corresponding linear X-ray ligand was used as the template for superposition. In cases of element mismatch in part of the structures, e.g. phenyl versus pyridine, between the linear ligand and its macrocyclic analogs, these atomic positions were still included in the RMSD calculations and are therefore marked in blue in Table 1. The RMSD values were calculated using the heavy atoms of the structural elements shared between the linear compounds and the macrocyclic analogs as annotated in Table 1. The corresponding linear X-ray ligand was used as the template for superposition. In cases of element mismatch in part of the structures, e.g. phenyl versus pyridine, between the linear ligand and its macrocyclic analogs, these atomic positions were still included in the RMSD calculations and are therefore marked in blue in Table 1. The RMSD values were calculated using the heavy atoms of the structural elements shared between the linear compounds and the macrocyclic analogs as annotated in Table 1. The corresponding linear X-ray ligand was used as the template for superposition. In cases of element mismatch in part of the structures, e.g. phenyl versus pyridine, between the linear ligand and its macrocyclic analogs, these atomic positions were still included in the RMSD calculations and are therefore marked in blue in Table 1.
Table 1
Structures of the linear compounds and the macrocyclic analogs. The name of the crystallized protein target, PDB code and, the PDB code of the corresponding linear structure is given in the top left, bottom left, and bottom right corner, respectively. The shared structure elements between the macrocycles and the linear structures used in the RMSD calculations are highlighted in blue.

(continued on next page)
3. Results and discussion

3.1. Data set

To facilitate comparisons between linear compounds and macrocyclic analogs, structure series were collected from the PDB where the linear compound and the macrocyclic analog were binding to the same protein, in a similar pose and making similar interactions with the target protein. In the data set, the linear ligands and the corresponding macrocyclic analogs are making the same hydrogen bond interactions with the protein target except for the Beta-secretase 1 (4DI2), HSP-90 alpha (4NH8), Coagulation factor XIa (4X6O) and Coagulation factor XIa (4Y8X) series. In these four series, all or some of the macrocycles had one additional hydrogen bond to the protein compared to their linear counterpart. In total, 39 structures consisting of 29 macrocycles and 10 linear compounds were selected from the PDB. The chemical structures and illustrations of the binding conformations for the 39 ligands are shown in Table 1 and Fig. 1, respectively. The data set comprises 10 compound series containing one linear ligand and at least one macrocyclic analog. The ligands are binding to 8 different protein targets, which means that two protein targets (Beta-secretase 1 and Coagulation factor VIIa) have multiple compound series co-crystalized. The target proteins and the PDB code of the corresponding linear ligand in this study were: Beta-secretase 1 (3IVH, 4DI2), Coagulation factor VIIa (4NGA), Heat shock protein (HSP) 90-alpha (4NH8), Coagulation factor XIa (4X6O, 4Y8X), HCV NS3/4A protease (5EQQ), ALK tyrosine kinase (5KZ0), Protein-tyrosine kinase 2-beta (5TOB) and Plasma kallikrein (6O1S). Henceforth the ligand series will be referred to by the target name. In cases where two series are targeting the same protein target, the PDB code of the crystallized linear ligand in each compound series will be added. To increase the size of the data set, we searched the original publications and were able to add 128 non-X-ray ligands (115 macrocycles and 13 linear analogs) to the ten series mentioned above. Thus, the complete data set comprises 167 ligands with 23 linear ligands and 144 macrocycles. The chemical structures for all ligands in the data set are shown in the Supporting Information.

3.1.1. Data set: properties

Due to the macrocyclic linker, the number of rotatable bonds is slightly higher for the macrocycles (8–23 rotatable bonds) compared to the linear compounds (4–18 rotatable bonds). The polar surface area (PSA) is also slightly higher for the macrocycles due to the addition of heteroatoms in the linker. However, most of the linkers in the data set are comprised of carbon atoms resulting in a higher calculated logP for the macrocycles compared to the linear compounds. The macrocyclic ring size varies from 11 to 20 with the majority (92%) of the macrocycles having a ring size of 15 ring atoms or less.

For the 39 X-ray ligands, the resolution varied between 1.41 and 2.39 Å and all ligands were consistent with the 2Fo–Fc electron density maps contoured at the 1σ level.

3.1.2. Data set: binding affinity

One of many reasons to prepare a macrocyclic analog is to improve its binding affinity over the linear analog. Hypothetically, since it can be expected that the macrocyclic conformational ensembles could be more shifted towards the bioactive conformation compared to the linear compounds, the binding affinities should be higher for the macrocycles. Thus, we collected activity data for the compounds to enable biological activity comparisons between the linear compounds and the macrocyclic analogs. Both IC_{50} and K_{i} data were collected. Out of the 167 ligands in the data set, binding and inhibition data were found for all ligands except for 6DIU. Furthermore, 4CLI, 4CTB, and 4CTC in the ALK tyrosine kinase series were excluded since only binding data for mutant target proteins were identified. In the HSP-90 alpha series, three ligands (denoted 4NH8_3R92_17, 4NH8_3QTF_33 and 4NH8_3QTF_34 in Supporting Information) were inactive. In this study, pIC_{50} and pK_{i} values will be collectively referred to as “binding affinity” data. All series had either pIC_{50} or pK_{i} data reported except HSP-90 alpha that had a mixture of values and was therefore removed from this analysis. Thus, 113 ligands were included in the binding affinity analysis. In Fig. 2, the differences in binding affinity between the corresponding linear ligand and its analogs is depicted. The binding affinity range was greater for the macrocycles (min: 5.2, max: 10.7) compared to linear compounds (min: 6.9, max: 8.6). However, as seen in Fig. 2, only four series (Coagulation...
factor VIIa, Coagulation factor XIa (4Y8X), ALK tyrosine kinase and, Protein-tyrosine kinase-2-beta) contain macrocyclic ligands with an improved binding affinity exceeding more than one order of magnitude. In fact the $\Delta$ binding affinity for most of the macrocyclic ligands varies by less than $\pm 1$ log unit as compared to their linear analog. Thus, it seems that the macrocyclization of the linear compounds in these series did not have a large impact on binding affinity. As mentioned before, to enable a comparison between macrocyclic and linear compounds, great care was deployed when collecting the compounds from the PDB. One criterion was to, as far as possible, make sure the macrocycles and the linear analogs were binding to the same protein target, in a comparable pose whilst making similar interactions. Out of the four mentioned series, only Coagulation factor XIa (4Y8X) contained macrocycles with one more hydrogen bond when compared to the linear counterpart. Although we have tried to select ligands that interact with the proteins in a very similar fashion, other factors may have effect on the binding affinity such as different substituent patterns.

3.2. Directing the conformational ensemble through macrocyclization

Macrocyclization can be used for many different reasons such as increasing bioavailability, metabolic stability, and binding affinity, removing off-target interactions, or generating IP. Regardless of the aims, the synthesized macrocyclic analogs need to fit into the protein binding site and be able to adopt the bioactive conformation of the linear counterpart to retain the biological activity. In the literature, macrocycles are often referred to as having a degree of conformational pre-organization, meaning that the conformational ensemble is less flexible and more shifted towards the bioactive conformation. Thus, the conformational distribution of the common structural element shared by

Fig. 1. Illustration of similarities between the X-ray conformations between the linear structure (gray) and the macrocyclic analogs (other colors). The protein structures were overlaid using all amino acid residues in the protein using the linear containing PDB structure as the template. To enable comparison between the X-ray ligands, the protein was hidden. The images are annotated with the name of the protein target and PDB code of the linear ligand in every series.
the macrocyclic analogs and the linear counterpart could be very different. To address this we performed a conformational analysis of the compounds in the data set and analyzed if the conformational ensembles of the macrocycles were more focused towards the bioactive conformation as compared to the linear structures.

3.2.1. Conformational distribution

The conformational distribution analysis was performed by calculating the RMSD of all conformers using the shared structural elements highlighted in Table 1 and the corresponding linear X-ray ligand as a template (see Supporting Information for the whole dataset). To create a reference point, we first analyzed the RMSD value for the X-ray conformation and the corresponding energy minimized X-ray conformation before analyzing the conformational ensembles. In the current study, a conformer was considered to be similar to the bioactive conformation if the RMSD value (using the shared elements) between the two conformers was below 1 Å.

For the 29 macrocycles available in the PDB, the RMSD between the linear X-ray ligand and the macrocyclic X-ray ligand using the shared elements varied between 0.18 and 0.38 Å (median RMSD equals 0.27 Å). Thus, the shared elements are binding in very similar conformations, see Fig. 1. Since the conformational sampling will only find local energy minima in a GB/SA water model environment, we also investigated possible differences between the energy minimized X-ray conformer and the corresponding linear protein-bound X-ray conformation. The energy minimized X-ray ligands had a median RMSD to the linear X-ray ligand of 0.51 Å (min: 0.22 Å, max: 1.72 Å, seven above 1 Å). This meant that large structural changes are taking place for some X-ray ligands when they are minimized outside of the binding site using a solvent model. Therefore, some of the X-ray ligands are not very close to a local minimum in an aqueous solution.

Using an energy window of 15 kcal mol\(^{-1}\), Fig. 3 shows how the conformational ensembles of all 167 ligands (linear and macroyclic) are distributed against their corresponding linear X-ray structures in terms of RMSD. The results suggest that the conformational distributions of the macrocycles were in most cases similar compared to those of the linear counterpart and not very focused towards the bioactive conformation. As seen in Fig. 3, mostly small shifts towards the bioactive conformation (i.e. RMSD equals 0 Å) are observed for the macrocyclic analogs. The RMSD values of the energy minimized linear X-ray ligands as compared to the X-ray ligand are marked with vertical gray lines in Fig. 3 (RMSD varies between 0.22 and 1.72 Å). As seen in Fig. 3, few macrocycles seem to have large portions of the conformational ensemble below the 1 Å RMSD threshold. Instead, the majority of conformers have an RMSD between 2 and 3 Å, which is not geometrically similar to the X-ray conformation. Nonetheless, in the ALK tyrosine kinase, HSP-90 alpha, and Protein-tyrosine kinase 2-beta series conformers very similar to the corresponding linear X-ray conformation (RMSD below 1 Å) were observed.

To summarize, as seen in Fig. 3, the conformational distributions of the linear ligand and the macrocyclic analogs are in most cases similar. However, the conformational distributions are affected by the macrocyclization since the RMSD values are not identical and the majority of the macrocyclic conformers are observed to be slightly more shifted towards the bioactive conformation. However, we argue that a macrocycle would be considered pre-organized only if the conformational ensemble has a high fraction of conformers similar to the bioactive conformation (RMSD below 1 Å). Using this classification, only three out of the ten series in the current data set contains pre-organized ligands. Thus, the conformational analysis of the ligands in the current data set does not suggest that the biologically active macrocycles are much more conformationally pre-organized into their bioactive conformation when compared to their closely related linear analogs. However, in a few cases, the macrocyclization did result in pre-organized ligands. The conformation distribution pattern did not change substantially using different energy windows (Supporting Information).

3.2.2. Conformational restriction

Due to macrocyclization, some conformations that are accessible in linear structures may become unavailable. In this section, we will investigate if the conformational space, measured as the number of conformers, was reduced for the macrocyclic analogs. Analyzing the number of generated conformers may, however, be misleading as the macrocyclic linker and the different substituents may adopt several conformations. Instead we will evaluate the number of conformers the linear ligands and their macrocyclic analogs are in most cases similar. To achieve this, a redundant conformational elimination was run using only the shared element (highlighted atoms in Table 1 and Supporting Information) thereby only saving conformers with a unique conformation in the shared element. Moreover, reaching conformational coverage for a conformational search using a 15 kcal mol\(^{-1}\) energy window for large ligands can need several hundred thousand search
steps. Thus, the 30,000 steps used in the current work are likely insufficient to reach full conformational coverage. However, since the number of possible conformers is lower for a more narrow energy window, a more complete conformational analysis can be expected using a smaller energy window. Therefore, in this section we only analyzed the number of generated conformers within 5 kcal mol\(^{-1}\) from the global energy minima.

As seen in Fig. 4, the number of unique conformations in the shared element within 5 kcal mol\(^{-1}\) of the global energy minima are generally lower for the macrocycles when compared to their linear counterparts. Thus, macrocyclization does seem to restrict the number of unique conformations the macrocyclic ligands can adopt within the shared

---

**Fig. 3.** Histograms (left) and cumulative plots (middle and right) visualizing the RMSD distribution for linear ligands (blue), and macrocyclic analogs (orange) towards the X-ray structure (RMSD equals 0 Å) using the atoms in the shared element and the corresponding linear ligand as a template. The dashed gray line indicates the RMSD of the corresponding linear ligand after energy minimization.
element but interestingly, not in all cases.

3.2.3. Linking the chemical structure to the conformational distribution and molecular flexibility

As shown in the cumulative RMSD plots (Fig. 3), the two Beta-secretase 1 (3IVH, 4DII2) series contain the macrocycles that are least affected by macrocyclization. This could be attributed to the large number of rotatable bonds outside of the macrocycle, which are no more restricted than the corresponding rotatable bonds in the linear analog (Table 1 and Supporting Information). Accordingly, as only a smaller part of the ligands are affected by the macrocyclization, and the remaining parts of the ligands remain highly flexible, the macrocyclic analogs will mimic the conformational properties of the corresponding linear ligand. In both series, this is confirmed by the number of unique conformations of the shared element where the macrocycles and the corresponding linear ligands are similar (Fig. 4).

The ligands in the Coagulation factor Xla (4Y8X) series are yet another example of only partly rigidified ligands where

Fig. 3. (continued).
macrocyclization only affects two of the rotatable bonds for the linear ligands (Table 1 and Supporting Information). Consequently, the macrocyclic conformational space is not largely affected by macrocyclization and is, therefore, more similar to the linear counterparts (Fig. 3). However, compared with the two Beta-secretase 1 series (3IVH, 4D12), a lower number of unique macrocycle conformations in the shared element is observed for the Coagulation factor Xla (4Y8X) series. This difference can be rationalized by the more rigid shared element of the Coagulation factor Xla (4Y8X) ligands.

Ligands of the two Coagulation factor Xla series investigated (4X6O and 4Y8X) displayed high structural similarity (Table 1). However, the ligands in the 4X6O series are slightly more rigid due to the six-membered ring between the macrocyclic part of the ligand and the phenyl chloride moiety. This explains why the Coagulation factor Xla (4X6O) macrocycles generally have slightly fewer unique conformations and also have somewhat more conformations focused on the bioactive conformation compared to the ligands in the Coagulation factor Xla (4Y8X) series.

Inhibitors in the HCV NS3/4A protease series also have large flexible substituents external to their macrocyclic motifs. However, a greater part of the HCV NS3/4A ligands are within the macrocyclic motif and therefore affected by the macrocyclization compared to the ligands in the Coagulation factor Xla. Hypothetically, the more rotatable bonds that are affected by the macrocyclization the higher dissimilarity to the

Fig. 3. (continued).

Fig. 4. The number of unique conformations in the common structural element using a 5 kcal mol\(^{-1}\) energy window. Linear ligands are depicted in blue and macrocyclic ligands in orange.
Macrocyclic ligands in the Plasma kallikrein binding series do not have large flexible substituents in the shared element, yet the conformational distribution is very similar to that of the corresponding linear ligand. This provides flexibility and, therefore, the conformational distribution does not differ significantly from the corresponding linear ligand. Conformational distributions within the Plasma kallikrein (6O1S) series are very similar (Fig. 3). However, the macrocyclization prohibits the macrocycles from adopting certain conformers, which could explain why the number of unique conformers in the shared element is lower for the macrocycles compared to the corresponding linear ligands within 5 kcal mol$^{-1}$.

Finally, the ligands binding to Coagulation factor VIIa, HSP-90 alpha, ALK tyrosine kinase and, Protein-tyrosine kinase 2-beta are among the compound series that in general have fewer conformers for the macrocycles compared to the linear ligand. The macrocycles in these four series are also among the few ligands that have conformers very similar to the X-ray conformation (below 1 Å RMSD) (Fig. 3). There are two common features for the macrocycles in these four series. First, a relatively large fraction of the ligands are affected by the macrocyclization (Fig. 1). Secondly, the macrocyclic ring size is rather small and contains fewer rotatable bonds without large flexible substituents. From our results, we can suggest that the following design features should be considered when targeting macrocycles that should populate the bioactive conformation: minimize the macrocyclic ring size whilst still maintaining a high percentage of the ligand within the macrocycle, and avoid the introduction of new rotatable bonds.

### 3.3. Evaluation of Boltzmann averaged RMSD approximating the conformational focus on the bioactive conformation

According to the Boltzmann distribution, different conformers will not be equally populated because of their energy differences. Consequently, increasing the population of conformers similar to the bioactive conformation should enhance protein binding. In 2018, Sindhikara and Borrelli presented Boltzmann averaging of RMSD values as a methodology for evaluating if potential macrocyclizations will focus the conformational ensemble on the bioactive conformation.$^{15}$ The Boltzmann averaging methodology has also been implemented in an automated macrocyclization tool.$^{15}$ In short, this methodology relies upon the correlation between binding affinity and the Boltzmann average of the RMSD values of the conserved elements (elements shared by the macrocyclic and linear ligands), denoted (RMSD). In this section, we evaluate (a) the difference in (RMSD) between the linear ligands and the macrocyclic analogs and (b) if the methodology can be used for prioritizing more potent compounds for synthesis. The Boltzmann averaged RMSD value (RMSD) was calculated for all ligands in the data set. As seen in Fig. 5, most macrocycles have lower (RMSD) values in comparison to their corresponding linear counterparts. However, only the HSP-90 alpha and ALK tyrosine kinase binding series contained ligands with (RMSD) values very close to the bioactive conformation (below 1 Å RMSD). All or nearly all ligands in the Beta secretase 1 (3IVH, 4D12), Factor Xla (4X6O, 4Y8X), HCV NS3/4A protease and, Plasma Kallikrein series have (RMSD) values above 2 Å, indicating an absence of pre-organization. Thus, when evaluated with the (RMSD) metric, macrocyclizations in most of the series of the current data set did not result in biologically relevant pre-organizations. However, in two cases (ALK tyrosine kinase, and Coagulation factor VIIa), the (RMSD) indicates that the conformational ensembles are focused on the bioactive conformation.

We next evaluated if the Boltzmann averaged RMSD metric could be used for prioritizing ligands for further analysis or synthesis. Out of the ten series investigated, only four series (Coagulation factor VIIa, Coagulation factor Xla (4Y8X) ALK tyrosine kinase and, Protein-tyrosine kinase 2-beta contain analogs at least one order magnitude more potent compared to the corresponding linear ligand (Fig. 2). Furthermore, some of the macrocycles in the Coagulation factor Xla (4Y8X) series have one more hydrogen bond to the protein compared to the corresponding linear ligand. Thus, the remaining analysis in this section will mainly focus on the Coagulation factor VIIa, ALK tyrosine kinase series and, Protein-tyrosine kinase 2-beta. However, since only three compound series will be analyzed, making general conclusions will be inappropriate. Nevertheless, as seen in Fig. 6, a clear correlation between binding affinity and RMSD can be observed only in the ALK tyrosine kinase series. This might be due to the very low (RMSD) values in the ALK tyrosine kinase series and the much higher (RMSD) values in the Coagulation factor VIIa series and, Protein-tyrosine kinase 2-beta. Thus,

![Fig. 5. The Boltzmann average of the RMSD values of the conserved elements (elements shared by the macrocyclic and linear ligands) is called (RMSD) for all ligands in the data set. Macrocycles are colored in orange whereas linear ligands are depicted in blue.](image-url)
the Boltzmann average method might only be applicable when \( \langle \text{RMSD} \rangle \) is not very high (i.e. 1–1.5 Å). Alternatively, the methodology could be used as an early filter removing ligands with a low probability of being pre-organized. However, the collected data set might be biased since all ligands most likely have passed through several stages in the medicinal chemistry workflow to ensure that the suggested macrocyclization will have a high probability of being biologically active and synthetically feasible. Furthermore, not all biologically inactive ligands might be included in the publications. Thus, we could not evaluate if the \( \langle \text{RMSD} \rangle \) methodology is an effective tool for removing unsuitable macrocyclization suggestions since no such suggestions are likely to be present in the collected data set.

To summarize, the methodology could show to be a useful tool for prioritizing macrocyclic analogs for further analysis or synthesis. However, due to the data origin and the small differences in binding affinity within most of the collected compound series, we were unable to fully evaluate the methodology. Thus, more analysis of the methodology with a more suitable data set is needed. Furthermore, even if the analogs were interacting similarly with the protein, other forces than the contribution of conformational changes such as desolvation will also affect the binding process. Thus, more advanced methods such as free energy perturbation calculations are needed to understand the differences in binding affinity. However, this is outside the scope of the current study.

3.4. Conformational energy window considerations

Many computational medicinal chemistry tools that involves conformational sampling require a conformational energy window to be specified. We therefore set out to investigate the energy differences between the global energy minima and the energy of the fully minimized X-ray conformations. Our results showed that the macrocycles and the linear compounds had a median energy difference of 4.6 kcal mol\(^{-1}\) and 4.8 kcal mol\(^{-1}\), respectively. Another way to find a suitable lower limit for the energy window is to evaluate the energy difference between the conformer closest to the X-ray ligand generated during the conformational search (i.e. conformer with lowest RMSD value) and the global energy minimum. For the 144 macrocycles in the data set, the macrocyclic conformer closest to the corresponding linear X-ray ligand had a median RMSD of 0.52 Å and a median energy difference to the lowest energy conformer of 6.5 kcal mol\(^{-1}\) (min: 0 kcal mol\(^{-1}\), max: 14.9 kcal mol\(^{-1}\)). This is in line with our previous work where the median energy difference between the lowest energy conformer and the conformer closest to the X-ray conformer (98% below 1 Å RMSD) was 6.6 kcal mol\(^{-1}\) (min: 0 kcal mol\(^{-1}\), max: 21.4 kcal mol\(^{-1}\))\(^{13}\). Taken together, this suggests that a more narrow energy window, compared to the commonly used energy window of 15 kcal mol\(^{-1}\), could be used in many cases, which can greatly reduce the number of conformations for consideration.

Fig. 6. Boltzmann averaged RMSD values (\( \langle \text{RMSD} \rangle \)) plotted against binding affinity for linear ligands (blue) and macrocyclic analogs (orange). (RMSD) above 2.0 Å are shaded in red.
4. Conclusions

Synthetic macrocyclic compounds are a privileged structural class in medicinal chemistry. However, they present numerous synthetic challenges and the field can greatly benefit from computational tools to help design new macrocycles more efficiently. Understanding how conformational preferences change for a linear ligand upon macrocyclization is central to this. To study this we collected a unique data set comprising protein structures from the PDB that contain both linear ligands and macrocyclic analogs, where each subset binds to the same protein, in a similar binding mode whilst making similar interactions with the target protein. This data set was expanded with synthesized ligands from the associated publications. In total, the collected data set comprised 167 ligands binding to 8 different protein targets. We believe that the collected data is a valuable resource for the community, and may be built upon and used to investigate, train, and validate other macrocycle-related computational tools and topics in the future. To facilitate the usage of the data set, data files that containing the X-ray ligands, X-ray proteins, energy minimized X-ray ligand structures, and the global energy minimum for each ligand are available in the Supporting Information.

Using the collected data set, we set out to investigate if the macrocyclic conformational ensemble is shifted towards conformers more similar to the bioactive conformation when compared to their linear counterparts. Interestingly, in many cases the macrocycle conformational ensemble distributions were not very different from those of the linear compounds. However, there are several other reasons for macrocyclization beyond pre-organization, such as improving bioavailability, toxicity, and other properties. According to the Boltzmann averaged metric (RMSD), only two out of the ten series in the data set were conformationally focused on the bioactive conformation and, on no strong correlations between (RMSD) and binding affinity were found. However, macrocycles did trend slightly towards favoring the X-ray conformation.

To summarize, our study shows that synthesized macrocycles obtained from the literature neither has to result in a higher binding affinity nor a higher degree of pre-organization, which might be of high relevance when designing new synthetic macrocycles. Consequently, macrocyclization does not necessarily have to lead to more preorganized ligands, not even in cases where the macrocycles are binding in a similar way as the corresponding linear ligands. However, in the collected data set, macrocycles with few rotatable bonds where most of the ligand is contained within the macroyclic motif had a higher chance of being focused towards the bioactive conformation. Thus, it is important to carefully design the macrocycle linker and the point of attachment to the linear molecule in order to focus the compound towards the bioactive conformation. Evaluations of the Boltzmann averaged methodology to approximate the conformational focusing on the bioactive conformation could be a useful tool for prioritizing macrocyclic analogs for further analysis or synthesis. However, we believe that a more thorough analysis of the methodology with a more unbiased data set is still needed. In terms of molecular flexibility, the number of unique conformers identified within 5 kcal mol⁻¹ of the global energy minimum was generally lower for the macrocycles compared to their linear counterparts. Finally, conformers close to the bioactive conformation were often found within 4.6–6.5 kcal mol⁻¹ from the lowest energy conformer generated. This suggests that a more narrow energy window, compared to the commonly used energy window of 15 kcal mol⁻¹, could be used for conformational searches in many cases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Swedish Research Council (521–2014-6711). The authors would like to thank Dr Lindon Moodie for constructive criticism of the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bmc.2021.116399.

References
