Journal of Computer-Aided Molecular Design (2021) 35:399–415
https://doi.org/10.1007/s10822-020-00337-7

ReSCoSS: a flexible quantum chemistry workflow identifying relevant solution conformers of drug-like molecules

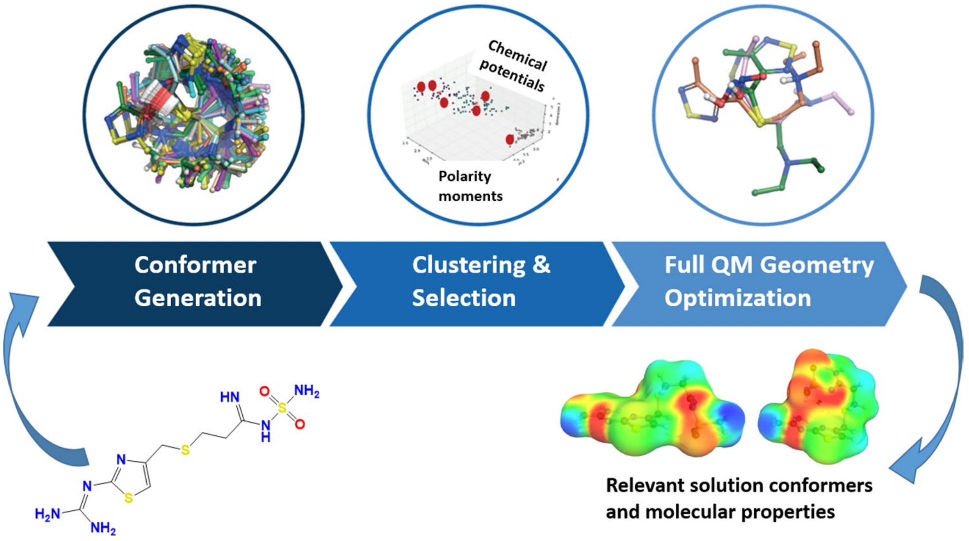
Anikó Udvarhelyi[1] · Stephane Rodde[2] · Rainer Wilcken[2]

Received: 4 June 2020 / Accepted: 12 August 2020 / Published online: 17 August 2020

Abstract

Conformational equilibria are at the heart of drug design, yet their energetic description is often hampered by the insufficient accuracy of low-cost methods. Here we present a flexible and semi-automatic workflow based on quantum chemistry, ReSCoSS, designed to identify relevant conformers and predict their equilibria across different solvent environments in the Conductor-like Screening Model for Real Solvents (COSMO-RS) framework. We demonstrate the utility and accuracy of the workflow through conformational case studies on several drug-like molecules from literature where relevant conformations are known. We further show that including ReSCoSS conformers significantly improves COSMO-RS based predictions of physicochemical properties over single-conformation approaches. ReSCoSS has found broad adoption in the in-house drug discovery and development work streams and has contributed to establishing quantum-chemistry methods as a strategic pillar in ligand discovery.

Graphic abstract



Keywords Conformations · Quantum chemistry · GFN-xTB · COSMO-RS · logP · ReSCoSS

Introduction

The prediction of low-energy conformers of drug-like organic molecules is a long-standing problem in computational chemistry. Designing ligands towards optimal binding to protein sites of different size and polarity, hydration

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-020-00337-7) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article patterns and plasticity, while simultaneously optimizing physicochemical and pharmacokinetic properties, is the essential challenge in drug design [1, 2] and relies in no small part on the correct description of tautomeric or protomeric forms, each featuring their own associated conformational space. Add to this the drug development challenge of polymorphism (ability of a molecule to crystallize in more than one crystal form) in crystalline solid dosage forms [3, 4] and the need to control potential conformational polymorphic transformations [5], it is clear that a method capable of accurately and routinely identifying relevant solution conformers of drug-like molecules would be of great value in the pharmaceutical industry.

Finding "relevant conformations" is a non-trivial task that may be roughly split into two parts: sampling the conformational space and ranking the conformers in terms of their relative free energy in a particular solvent or set of solvents. Keeping in mind that conformational flexibility is an important aspect to ligand binding, it is not surprising that conformer generation has been studied extensively in the field of computer-aided drug design [6, 7]. Recent assessments of available open-source tools as well as commercial algorithms have demonstrated somewhat converging performance of the most widely used generators in recapitulating protein-bound ligand conformations among the first few hundred conformers generated [8, 9]. With a selection of decent conformer generators to choose from, we focused our efforts on the second step: ranking the conformer energies in solution. Since conformer weights depend exponentially on the relative free energies (as described by the Boltzmann distribution), relatively small errors in calculated energies can result in large predicted shifts of the conformer equilibrium. Most papers studying the performance of conformer generators do not in fact discuss conformer energetics but rather focus on the ability of the algorithms to recapitulate known protein-bound conformations from crystal structures within a certain RMSD threshold (e.g. [8, 10]). In manuscripts where there is a focus on energetics, unsurprisingly, most current small-molecule force fields are shown not to be accurate enough [11, 12]. A very recent exhaustive study spanning force fields, machine-learning potentials, semiempirical methods, wave-function methods and density functional theory suggests that calculations at dispersion-corrected DFT level may provide a good compromise between speed and accuracy [13]. This is particularly crucial in the fast-paced pharmaceutical industry environment. Modelling solvation effects accurately is another important piece of the puzzle for use in drug-design applications. With explicit solvation methods being prohibitively expensive at the DFT level, we chose to model solvent effects using COSMO-RS [14–16] which has been shown to afford good performance when combined with DFT-D for the calculation of host–guest complexes [17] and also allows the calculation of physicochemical properties like partition coefficients and ionization constants [18–21].

Drug-like molecules do not necessarily bind their biological targets in a minimum-energy conformation [22, 23] so it is important to conduct thorough conformational sampling. Considering the rather inaccurate energies of small-molecule force fields that are used to drive the conformer generators, this often means having to postprocess hundreds of conformations per compound of interest in a refinement step with more accurate quantum chemical methods including solvation. The COSMOconf workflow, for instance, employs a hierarchical scheme starting from force field geometries and energies and then conducts energy assessments and geometry optimizations at increasingly involved levels of theory [24, 25]. Depending on the thresholds set in such a workflow (for example, number of conformers to be optimized at the highest level of theory), the resulting calculation times are rather long; on the other hand, setting too stringent cut-off values on the number of conformers to be considered during the workflow runs the risk of discarding relevant conformers that have higher energies at lower level of theory. We therefore set out to design a different workflow to allow the identification of only a handful of diverse and representative solution conformers per molecule without setting strict cut-offs on numbers and to only optimize those relevant low-energy conformations with full DFT-D. Doing this required the introduction of a novel clustering and selection strategy employed after conformer generation. In accordance with its purpose, the workflow is dubbed Relevant Solution Conformer Sampling and Selection (ReSCoSS) and has found considerable adoption in our in-house discovery and development work streams.

ReSCoSS workflow architecture

We aimed to design a flexible and semi-automatic workflow that selects a relevant subset of conformers from a large conformer ensemble for full geometry optimization at DFT level and subsequent COSMO-RS energy calculations. A graphical overview is presented in Fig. 1a; further computational details are given at the end of the manuscript. ReSCoSS is a Python workflow that can use 2D SDF as input, which allows the user to sketch the molecule of interest in tools like ChemDraw or alternative sources of 2D coordinates. In a first step, the molecule set is split by tautomers (if tautomers exist) and initial 3D coordinates are generated using CORINA [26–28]. The next step, conformer generation, is interfaced with three tools, Schrodinger's Macromodel [29], CCG's MOE LowModeMD [30, 31], and RDKit ETKDG [32, 33] as open-source alternative, and these tools can be used on their own or in combination. ReSCoSS also allows inputting a user-defined set of conformers in case that is

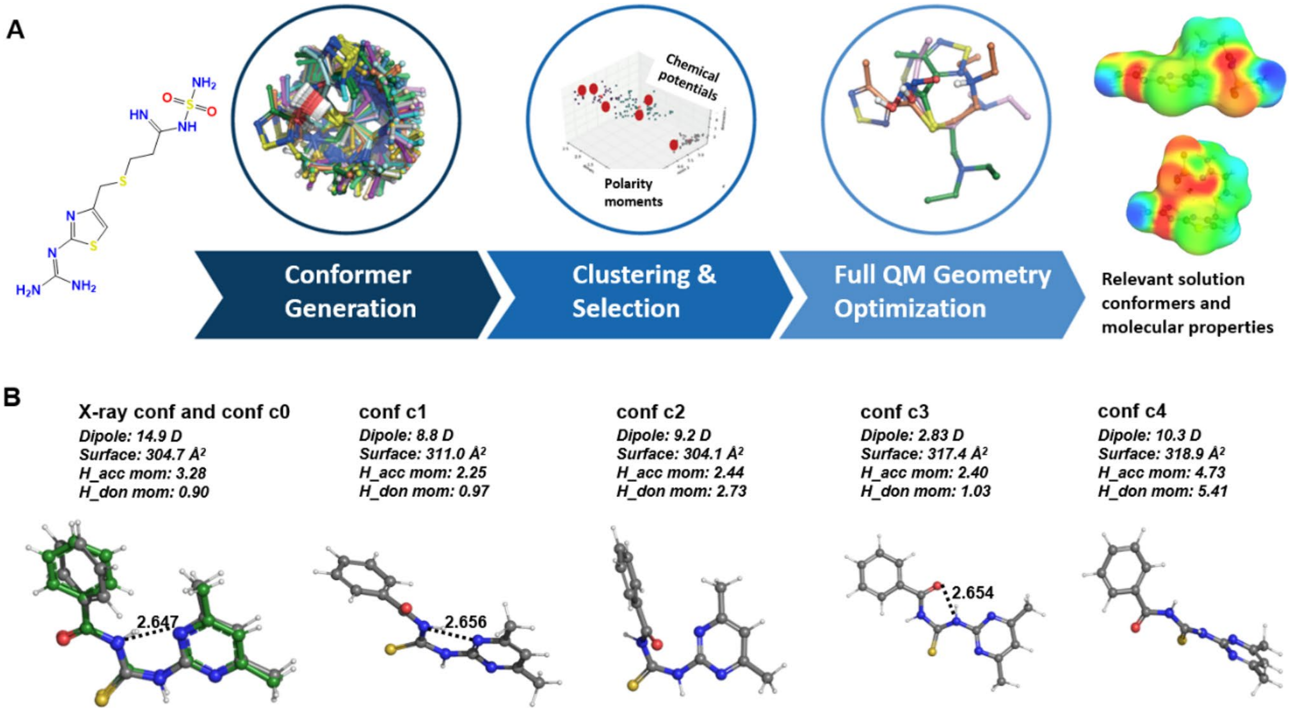


Fig. 1 The ReSCoSS workflow. **a** Graphical scheme highlighting the main steps in the workflow. **b** A closer look at the descriptors used in clustering using N-(((4,6-dimethylpyrimidin-2-yl)amino)carbonothioyl)benzamide (CSD: AWUBID) as an example molecule. Conformers shown include the crystallographic conformation (green carbons) overlaid onto the lowest-energy conformer (c0) from the conformer set (grey carbons) alongside other conformers. For each conformer, the values of the four descriptors used in the k-means clustering are shown

desired. This flexibility has proven to be quite useful in practice as using multiple tools can avoid sampling issues introduced by any particular conformer generator. The set of resulting conformers is handled as a 3D multi-SD file within ReSCoSS and in the next step all conformers of this set are optimized using GFN2-xTB [34] using the GBSA-water model with runtimes in the order of 1–2 s per conformer for a drug-like molecule.

The second step, clustering and reduction, constitutes the heart of the workflow and sets ReSCoSS apart from other available tools. We noticed that for relatively large drug-like molecules the usual hierarchical strategy of retaining large sets of low-energy conformers at low level of theory (e.g. force-fields or semiempirics methods), then reducing in further optimization steps and finally running more sophisticated calculations on the remaining set is often tedious and computationally expensive. However, reducing the size of the conformer set too early can be dangerous as well since it runs the risk of discarding relevant conformers early on and ending up with a non-representative set of conformers in the output. In some cases, especially when calculating ionized or very polar compounds, the limited accuracy in the description of solvation effects with fast models such as GBSA can also lead to problems. With ReSCoSS we aimed to tackle both issues: we introduce COSMO-RS solvation early in the workflow, expecting that this will lead to a more accurate energy description of the conformations; and we use a clustering scheme based on quantum-derived descriptors to define diverse shape classes out of which low-lying conformers are picked for further processing. The aim was to ensure that the selected subset to be used further for DFT-D optimization is diverse, so that it contains “extended” and “folded” conformations, those with and without intramolecular H-bonds, both “cis” and “trans” conformations, for example.

After the conformer set is fully optimized using the fast GFN2-xTB/GBSA method, we conduct single-point energy calculations for each conformer at the B97-3c/COSMO [35, 36] level of theory using Turbomole [37, 38]. B97-3c [35] is a relatively recent refit of the B97-D functional combined with a reduced triple-ζ basis (mTZVP) and strikes an ideal balance between speed and accuracy [13]. These single-point calculations take about two minutes per conformer. As we generate COSMO files for each conformer at this level, next to standard descriptors used for clustering such as RMSD we also evaluated descriptors derived from the σ-surfaces in the COSMO-RS framework [14–16] to characterize conformer shapes, as opposed to simple pure-geometry descriptors. We tested several different descriptor combinations and found that the dipole moment, the sigma

H-bond donor moment, the sigma H-bond acceptor moment and the sigma total surface area (all four calculated after the B97-3c/COSMO single point calculations from the COSMO output file) were the four most relevant descriptors. In combination, these four could well capture and distinguish the typical conformer shapes of drug-like molecules.

Using *N*-(((4,6-dimethylpyrimidin-2-yl)amino)carbonothioyl)benzamide (CSD: AWUBID) [39] as an example molecule, we show in Fig. 1b how well these descriptors can differentiate five different classes of conformations that contain different intramolecular H-bonding patterns and extended vs folded shapes. Conformers c0 and c1 exhibit the same intramolecular H-bond but differ in how extended they are, which is captured and differentiated by the sigma surface descriptor. Conformer c3 also exhibits an intramolecular H-bond, however, a different one involving the carbonyl, which causes a change in the molecular dipole moment. Conformers c2 and c4 do not have intramolecular H-bonds and hence have larger sigma H-bond acceptor and donor moments than the other conformations featuring an intramolecular H-bond. Conformers c2 and c4 are distinguished with a different COSMO surface area as the former is quite folded in shape and the latter completely extended.

Using these four COSMO-derived shape descriptors as features, we use the *k*-means clustering algorithm as implemented in scikit-learn [40] in Python to cluster the conformer ensemble in a four-dimensional shape space. The parameter *k* determines the number of clusters to be used in the clustering procedure. *k* is read as an input variable in ReSCoSS and it is not straightforward to establish its ideal value. Usual procedures in machine-learning applications to determine *k*, like the so-called elbow method or the silhouette method, are less suitable in our case as the shape-space clustering is just one part of the conformer selection procedure. We need to add the conformer energies and select the low-energy conformers from each shape cluster, rather than some representative average conformer from each cluster.

As we are interested in relevant solution conformers, it is decisive that we use solution energies during the conformer selection. We use the chemical potential, calculated by COSMOtherm [41, 42] (at the TZVP level) in ten different media: water, DMSO, cyclohexane, 1-octanol, methanol, chloroform, acetone, perfluoropyrrole, acetonitrile and vacuum. These solvents are chosen to ensure maximum coverage of possible dielectrics and H-bonding capabilities of the solvents. The example of AWUBID conformers in Table 1 shows that although conformer c0 is identified as the lowest-lying conformer in all solvents, the second-best ranked conformer varies between the solvents. Note that there are significant conformer free energy differences between different COSMO-RS solvents and also compared to the electronic energy from B97-3c/COSMO alone; this ability of COSMO-RS to differentiate clearly between the solvents by taking into account more than just the dielectric environment sets it apart from simpler implicit solvent models.

Having partitioned the whole conformer set into *k* clusters using the *k*-means procedure described above, we select *N* conformers from each cluster by their COSMO-RS chemical potential in each of the solvents. The two parameters *k* and *N* thus determine the conformer selection out of the total ensemble. Setting k = 1 corresponds to bypassing the shape clustering and simply selecting conformers based on their chemical potentials in the ten solvents. By varying the *k* and *N* combinations, denoted kN set in the following for simplicity, we can test how the conformer selection changes if we place more emphasis on shape diversity (by increasing *k*) or more emphasis on the COSMO-RS free energy landscape (increasing *N*). We expect that the ideal combination of *k* and *N* is dependent on the molecule of interest. Figure 2 visualizes the clustering step for a highly flexible drug, the histamine H2 receptor antagonist Famotidine [43]. In this case, the first step of the workflow (combined MOE and Macromodel conformer search) generates a conformer set of 404 conformers, featuring a large diversity of shapes. The clustering step for the case of *k* = 5 clusters is shown in Fig. 2a, where we show only two out of the four descriptor dimensions for visualization purposes (dipole moment and COSMO surface) together with the computed free energies in water.

In order to compare different kN combinations we use violin plots to visualize the spread in the descriptors. These are essentially one-dimensional histograms and allow easy comparison of different distributions. Figure 2b shows violin plots of the spread in the H-bond donor and acceptor moment descriptors in the full famotidine conformer set and

Table 1 Energetics of N-(((4,6-dimethylpyrimidin-2-yl)amino)carbonothioyl)benzamide (AWUBID) conformers

AWUBID Conf. #	ΔE (B97-3c/COSMO) (kcal/mol)	ΔG (COSMO-RS) in solvent (kcal/mol)			
		Acetonitrile	Chloroform	Water	Methanol
c0	0.00	0.00	0.00	0.00	0.00
c1	2.77	1.97	2.24	2.80	1.81
c2	5.70	2.56	4.38	2.36	1.96
c3	1.58	2.07	2.40	2.64	1.52
c4	8.52	6.48	10.94	4.37	5.55

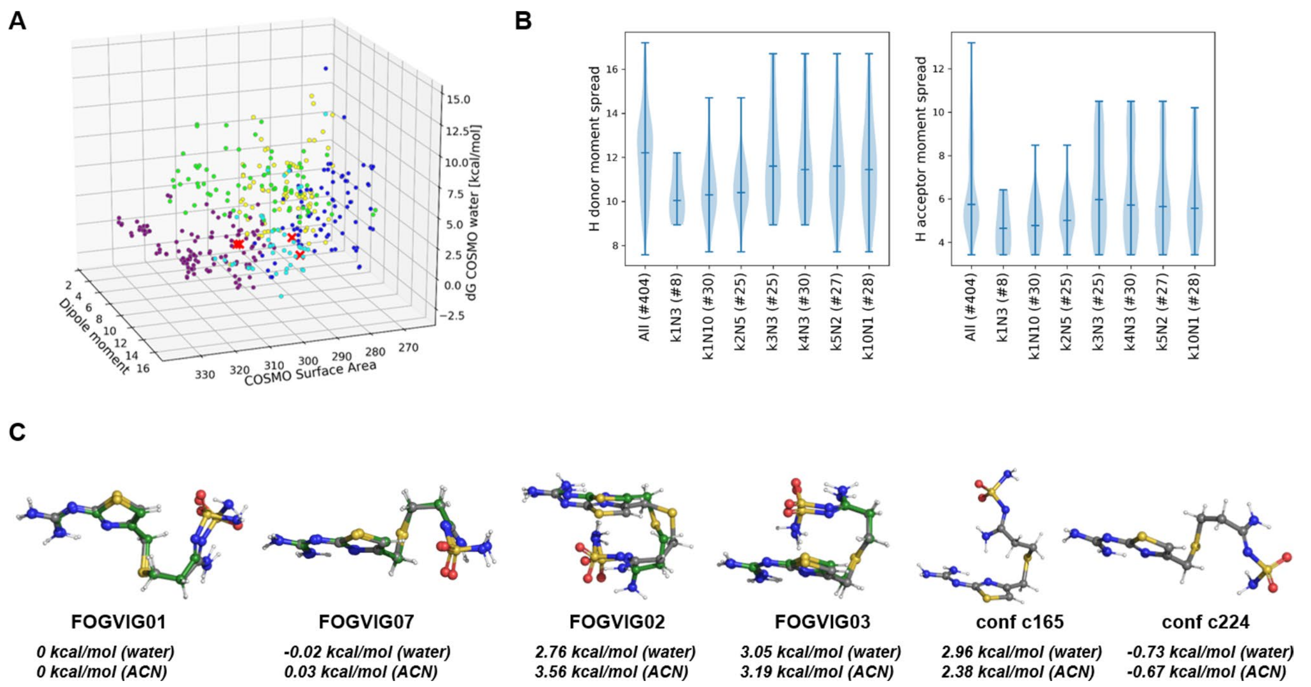


Fig. 2 Conformer clustering 404 conformers of Famotidine. **a** 3D scatter plot showing the dipole moment, COSMO surface area, and ΔG (COSMO-RS water) energies in water for k = 5. The different conformer clusters are represented by different coloring; red crosses indicate the experimentally known solid-state conformers. **b** Violin plots of the H-bond donor and acceptor moment descriptor spreads in the full conformer set and for selected kN combinations. The numbers in brackets indicate the number of conformers chosen for full optimization in each set. **c** Relevant famotidine conformations: four conformations from the CSD (green carbons) overlaid onto the most similar ReSCoSS conformer from the k1N10 set (grey carbons). Conformers #165 and #224 are the lowest-energy conformers in water at the selection step and final COSMO-RS energy assessment after B97-3c/COSMO optimization, respectively. Energies given are at COSMO-RS FINE19 level relative to the FOGVIG01 conformer, for water and acetonitrile

in subsets of different kN. The numbers in brackets indicate the total number of conformers in each set. For example, the notation k1N3 (#8) indicates a conformer set where no shape clustering was performed (k = 1) and the three lowest conformers were selected in any of ten solvents (N = 3), in this case resulting in eight selected conformers. Although we use ten different COSMO-RS solvents in the selection and pick the lowest-lying three conformers in each solvent (N = 3), significantly less than the maximum of 30 conformers are chosen in this case. In general, the relative ranking in terms of free energies among the solvents is often convergent for drug-like molecules, with one class of conformer being dominant in the more polar, hydrogen-bonding solvents such as water and methanol and another considered dominant in the apolar environment simulated by chloroform, cyclohexane or vacuum/gas phase. Once a conformer has been picked in one solvent environment, it is not picked again even if it corresponds to the *N* lowest-lying conformers in other solvents in order to avoid duplication.

From the plots in Fig. 2b it is clear that the k1N3 conformers have a small diversity in the H-bond donor and acceptor moments compared to the full set. With larger *k*, as expected, we obtain more diverse sets, which are close to reproducing the diversity seen in the whole conformer set.

The third step of the ReSCoSS workflow is simply a full geometry optimization of all conformers in the selected kN set at the B97-3c/COSMO level. Because this step is rate-limiting in terms of the total runtime of the workflow (taking up to 6 h per conformer), we tested several GGA and mGGA functionals (BLYP, BP86, PBE, TPSS) in combination with dispersion corrections and double- and triple-ζ basis sets and found that B97-3c was the best compromise in terms of accuracy and speed. Finally, a single point at the COSMO-RS FINE19 level (BP86-D3(BJ)/TZVPD/COSMO-FINE) is conducted. This is needed as input for COSMOtherm evaluations at the FINE19 level. Using COSMOtherm, calculation of Boltzmann weights in different solvents (chosen by the user) is conducted and various properties are calculated if desired, for example logP(o/w).

For Famotidine both the k1N10 and the k5N2 set contain three of the four known crystallographic conformations, as given in the CSD with refcodes FOGVIG01 and FOGVIG07 (extended conformations) as well as FOGVIG02 and FOGVIG03 (folded conformations), shown in Fig. 2c. At the selection step, the four experimental crystallographic conformations are located in three different clusters (marked with red crosses in Fig. 2a). Conformer c165 is the lowest energy conformer in water at the COSMO-RS (TZVP) level of theory before full optimization, however it is considerably less favored in energy after the full optimization (see Fig. 2c) and conformer c224 is the overall lowest in the set. This observation supports our strategy of not solely relying on calculated free energies during the workflow but combining it with shape diversity as well.

Assessing *k* and *N* through conformational case studies

Having established the principle of our conformer clustering and selection strategy, we next analyze the performance of ReSCoSS in detail on three diverse molecules as a function of the kN clustering. To this end, we fully optimized all conformers of the molecules at the B97-3c/COSMO level followed by COSMO-RS FINE19 single point energy calculations. This allows us to assess how well our selection strategy is capable of picking the relevant conformers and the lowest-energy conformers in solution. For all three molecules there are small-molecule single crystal X-ray structures available and crucially, for two out of three there is solution NMR data.

Analyzing conformational preferences of small drug-like molecules in the solid state, e.g. from crystallographic databases like the CSD [44], can provide an understanding of ground-state conformations and is often used to drive design strategies [45]. While these analyses are undoubtedly of great utility and have the advantage of relying on experimental data, they reflect behavior in the solid state and not necessarily conformations in solution. For example, it is known that molecules can crystallize in rather strained conformations if favorable packing interactions such as hydrogen bonding in the crystal lattice are involved. Molecules can also crystallize in two or more very different conformations showing large free energy difference as computed on the isolated conformers [4]. Accordingly, throughout the development of ReSCoSS, we used available crystal structures as reference structures that we aim to reproduce in the chosen kN subsets as they are clearly relevant but kept in mind that crystal structures do not always reflect the solution-state minimum.

One such example where the experimental crystallographic conformer exhibits a rather unusual *cis* carbamate conformation is the dipeptide Boc-Phe-m-aminobenzoic acid, CSD refcode YASNUD [46]. Other related dipeptide molecules with the aminobenzoic acid moiety in the para or ortho position crystallize in the more usual trans carbamate conformation, as observed in CSD entries TIFJAV and TIFJEZ. Out of the 334 conformers generated in the initial step for YASNUD, the solid-state cis carbamate conformation is ranked #52 at the selection step and clearly $k > 1$ clusters are needed to select and carry over the experimental conformation into the geometry optimization step. Figure 3 shows the relevant crystallographic and low-energy conformers, as well as the final energy distribution in different kN sets, visualized with violin plots. Only the k4N3 and the k10N1 sets include the crystallographic conformation. The best *cis* and best *trans* carbamate conformations in the crystallization solvent methanol are completely extended ones. The lowest-energy conformer in methanol in the final set is only picked by the k10N1 set. Importantly, the clustering results in a clear and significant reduction by more than 2 kcal/mol of the median energy in each kN set, compared to the full set. Notably, the k10N1 set (selection focusing

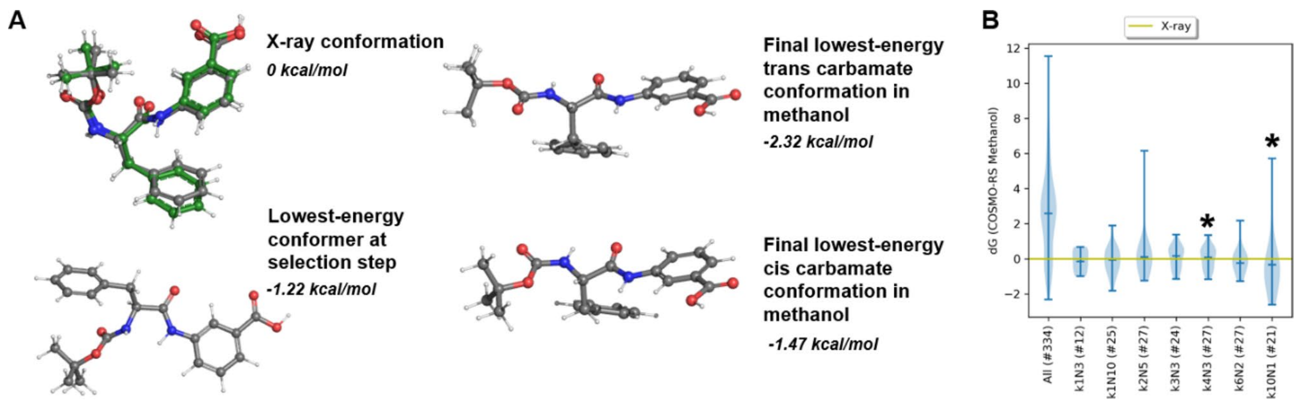


Fig. 3 Relevant conformations of YASNUD [46]. **a** Crystal structure (green carbons) overlaid onto the most similar structure from the conformer set after B97-3c/COSMO optimization (grey carbons). In addition, the lowest-energy conformers at the selection and final steps are also shown, including their COSMO-RS FINE19 energies in methanol, relative to the crystallographic conformer. **b** Final energy spread in methanol at the COSMO-RS FINE19 level for different kN subsets. The energy of the crystallographic conformation is highlighted with a yellow horizontal line. The asterisks indicate the sets that selected the exact crystal conformation

on shape diversity at the selection step) even has a lower median energy at the final energy level than the k1N10 set (selection only based on energies at the selection step). This indicates that the shape clustering does not introduce unrealistic high-energy conformations into the subsets but indeed results in low-energy diverse conformer sets.

Next, we looked at Compound 28 from [47] which describes a series of α-methylpiperidine carboxamide dual orexin receptor antagonists (DORA). In the absence of protein crystal structure for the orexin receptor, elucidating the conformational behavior of active ligands proved to be paramount and the minimization of 1,3-allylic strain was an important design principle. For DORA-Cpd28, a solution NMR structure in chloroform is described as reference in addition to a small-molecule X-ray crystal structure, CSD refcode VATBID. The two conformations are distinct as shown in Fig. 4a.

For DORA-Cpd28, the ReSCoSS full conformer set consisted of 216 conformations. Figure 4b shows the overlays of the selected conformers in different sets of kN. As expected, the k1N3 set results in a set of low diversity where several conformers are very similar. Picking $N = 10$ low-energy conformers at the selection step without shape clustering (k1N10) results in a selection of 17 out of the 216 conformers. In contrast, with the k3N3 selection strategy 16 conformers are selected yet the shape diversity is clearly higher than in the k1N10 case and the median energy of the set is similar in both cases as evidenced by the energy violin plots in Fig. 4c. In this case, we chose two different solvents for analysis, water and chloroform, the latter being relevant for the NMR conformation as chloroform was the solvent used in the NMR experiments. Again, the crystallographic conformation is used as a reference for the relative conformer free energies in each set and is indicated by the yellow horizontal line in the plots. There is another X-ray-like conformer in the set, indicated by the red horizontal line, featuring the same shape except for a rotated fluoropyridine ring, which lies 0.2 kcal/mol and 1.7 kcal/mol above the X-ray conformer in water and chloroform, respectively. The exact X-ray conformer is only picked in the k1N10 and k3N3 sets with k3N3 being most attractive from a diversity and subset conformer numbers standpoint.

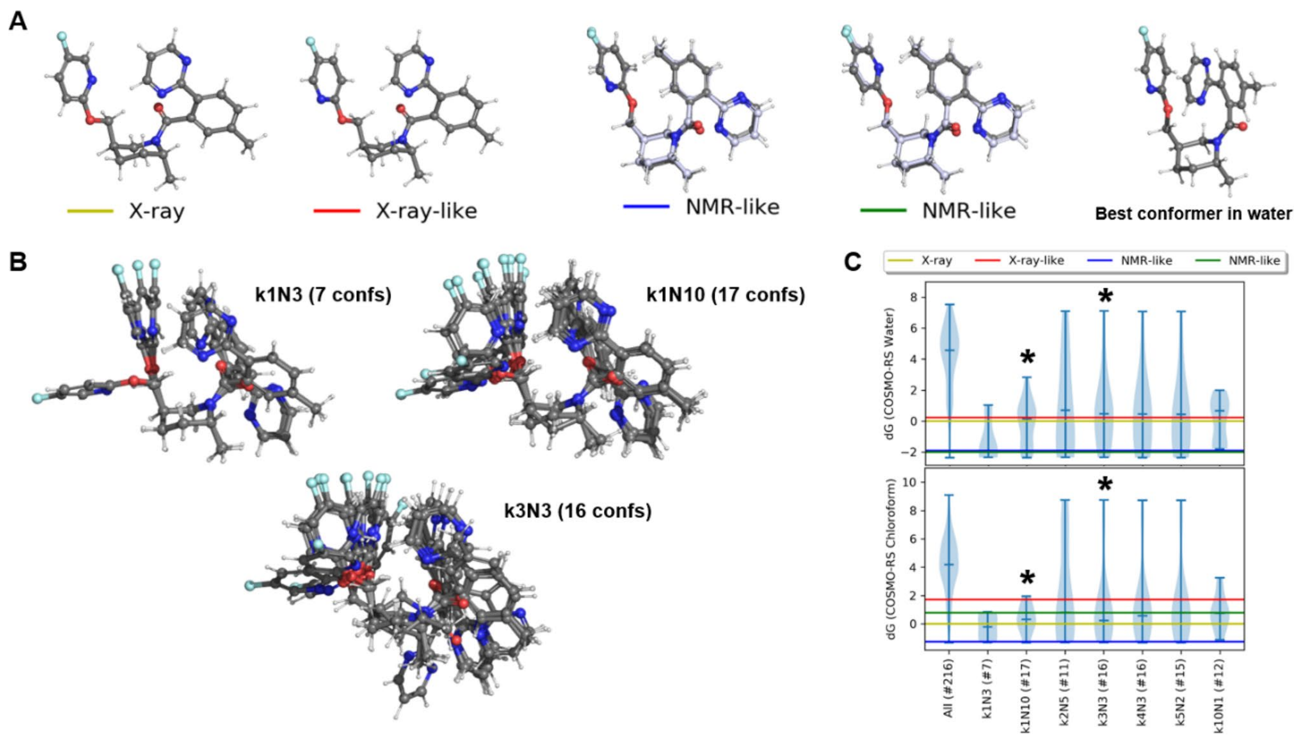


Fig. 4 ReSCoSS analysis of dual orexin receptor antagonist Cpd28. **a** Representative conformers from solid state (CSD refcode VATBID) and those fulfilling the ROE restraints from solution NMR [47] as well as the lowest-lying conformer in water. The conformers are all overlaid on the central piperidine moiety. **b** Overlay of the selected conformer subsets for the k1N3, k1N10 and k3N3 combinations, showing that k3N3 selects a more diverse set than k1N10 while the overall number of conformers is similar (16 vs. 17). **c** Relative free energy spread according to COSMO-RS FINE19 in water and in chloroform for the different kN combinations. In brackets the number of conformations in each set is given. The free energy of the X-ray conformation is used as reference (yellow line at y = 0 kcal/mol); the conformer with rotated fluoropyridine group is denoted "X-ray-like" and its energy is indicated with the red horizontal line. Two conformations fulfilling the NMR constraints are marked in blue and green. The sets annotated with an asterisk indicate sets that contain both conformers known from experiment

The alternative "X-ray-like" conformer is picked in all sets with k = 3, 4, or 5. Larger values of *k* do not select any of the two X-ray-like conformers. In addition, the k10N1 set also does not pick the best conformer in water that exhibits a stacking interaction between the fluoropyridine and pyrimidine rings. Two very similar conformers that differ by a rotation of the fluoropyridine moiety both satisfy the NMR constraints and are indicated by blue and green horizontal lines in Fig. 4. Interestingly, these two related conformations are quasi equienergetic in water and approximately 2 kcal/mol more favorable than the X-ray conformation. However, in chloroform one (blue) is more favorable than the other by about 2.1 kcal/mol, most likely because the rotation of the pyridine places the polar nitrogen in a less shielded position in the less favored conformer. The more favorable conformer in line with NMR constraints is picked in all tested kN sets, as is the alternative NMR conformation. It is very encouraging that ReSCoSS predicts the best conformer in chloroform correctly in line with the NMR results from [47].

In the final example, we studied a charged molecule, Compound 10 from a series of antagonists of X-linked inhibitor of apoptosis proteins (XIAP) [48], denoted XIAP-Cpd10 in the following, to assess the performance of ReSCoSS for non-neutral species. There is solution NMR data and an X-ray crystal structure of Cpd10 bound to XIAP-BIR3 available as reference (PDB: 5M6E). The binding pose in the protein pocket requires a folded conformation of XIAP-Cpd10 featuring intramolecular π-stacking between the pyrazole and amide moieties. In NMR studies in a phosphate buffer, ROEs consistent with a folded conformation were observed, leading the authors to conclude that the bound conformation is also the dominant conformation in water. We observe little dependence of the overall performance on varying *k* and *N*. All our kN sets picked the crystallographic conformation with the exception of k10N1. The energy violin plots among the different low- and mid- *k* sets are quite similar, as shown in Fig. 5b. Although the crystallographic conformation is indeed a low-energy conformation, we find that according to COSMO-RS, it is not the minimum-energy conformation in water environment as reported in [48]. We find that an additional similar but further folded conformation, shown in Fig. 5a, with the phenyl ring rotated towards the 4-methyl pyrazole, lies 2.2 kcal/mol lower in energy. Indeed, this conformation is also consistent with the distances observed in the ROESY experiment from [48] and might even explain the observed data better since the relevant H1-H7 and H2-H8 distances are shorter for our predicted conformation compared to the X-ray binding pose.

Although the ideal choice of *k* and *N* is not always the same and does depend on the molecule studied and its conformational landscape, the case studies shown here demonstrate two trends, namely that a very low *k* (e.g. 1 or 2) is not ideal because it usually leads to very uniform-looking conformer subsets while choosing a high value for *k* with associated small *N* leads to diverse sets as expected. In some cases, however, relevant conformers that do not correspond to the lowest-energy conformations at the clustering step due to less accurate energetics are missed out in sets of high *k* and low *N*. In our in-house applications we usually employ a medium *k* value and *N* > 1 in combinations such as k4N2, k5N2 and the balanced k3N3 which delivers medium-sized diverse conformer sets for optimization and has emerged as the de facto standard. Judging from the cases discussed here as well as several years experience in in-house application of ReSCoSS, we conclude that while there is certainly no one-size-fits-all k/N combination, the risk of discarding ultimately relevant conformers from the set at the selection step can be minimized by employing medium *k* and *N* > 1. In our experience, the inability of conformer algorithms to generate all relevant conformations in the first place is often the bigger issue than finding the ideal *k*/*N*. This can be mitigated by

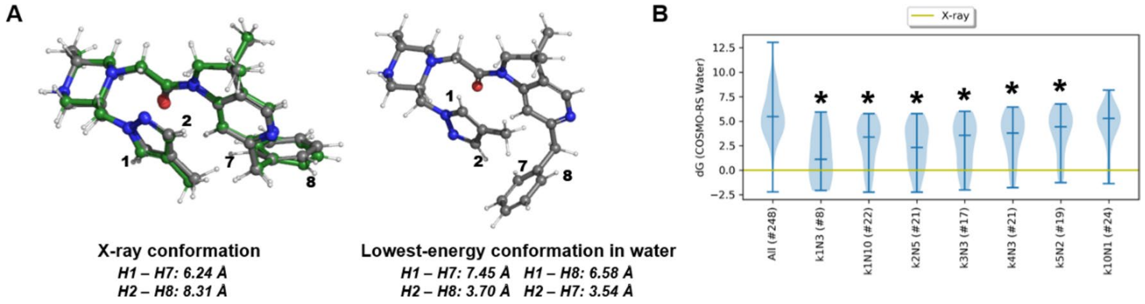


Fig. 5 ReSCoSS analysis of XIAP-Cpd10. **a** Left: Crystallographic binding pose (PDB: 5M6E, green carbons, from [48]) overlaid onto closest ReSCoSS conformer pose (grey carbons). Right: Minimum-energy conformation in water according to COSMO-RS FINE19. The hydrogen atoms that are relevant for the ROEs reported in [48] are indicated with the same numbering as in the original publication. **b** Relative free energy spread of the different kN subsets indicated with violin plots. The asterisks indicate sets where the bound conformation was picked during selection

using a combination of conformer generation algorithms and tools that ReSCoSS allows for.

Applying ReSCoSS in the industry environment

One of the central concepts in the development of the ReSCoSS workflow is to enable a more informed approach to molecular design by putting quantum-chemistry methods at the fingertips of the medicinal chemistry community. Due to the clustering and selection strategy described previously, only a limited number of relevant conformers for each molecular species have to be fully optimized at DFT level, which makes the workflow performant enough to enable turnaround times of hours to roughly a day for a molecule of interest at maximum in most cases. It is therefore routinely applied in-house in both structure-based and ligand-based design as well as in conformer generation for properties prediction in drug development. ReSCoSS provides a final set of conformers with COSMO-RS free energies and Boltzmann weights in a range of solvents chosen by the user. The ReSCoSS conformers serve as input for further property predictions using COSMO-RS such as free energy of solvation, partition coefficients such as logP(o/w) and pKa values, if protomers are considered and calculated in addition.

AZD5991: ReSCoSS applied to macrocycles

AZD5991 is a macrocyclic Mcl-1 inhibitor with selectivity over Bcl-2 and is currently in clinical studies for relapsed or refractory hematologic malignancies [49]. Its rational design was driven by AstraZeneca's conformational analysis platform which uses selected hydrogen positions in a molecule as "conformational ^{1}H NMR reporters" that allow to judge conformational preorganization in solution [50]. AZD5991 is a large (MW = 672.3 Da) and conformationally flexible molecule and hence represents a challenging test case for ReSCoSS. Of note, the authors remark in the original paper that the bioactive conformation is not the same as the calculated global minimum-energy conformation in water at the level of theory used (B3LYP/6-31G* with PCM-water) [50]; yet the bioactive conformation was experimentally confirmed as the dominant one in solution both in water and DMSO$_{\text{-d6}}$ as by NMR spectroscopy. ReSCoSS produces a total of 316 conformations, and after B97-3c/COSMO single points and k-means clustering using k = 3, N = 3, 20 diverse conformers are selected for full optimization (Fig. 6a). Gratifyingly, the conformation out of the final set of 20 corresponding to the global energy minimum in both water and DMSO according to COSMO-RS FINE19 is very similar the published bioactive and co-crystallized conformation (PDB: 6FS0) with a heavy-atom RMSD of 0.6 Å—the closest in the set. The aforementioned alternative conformation discussed

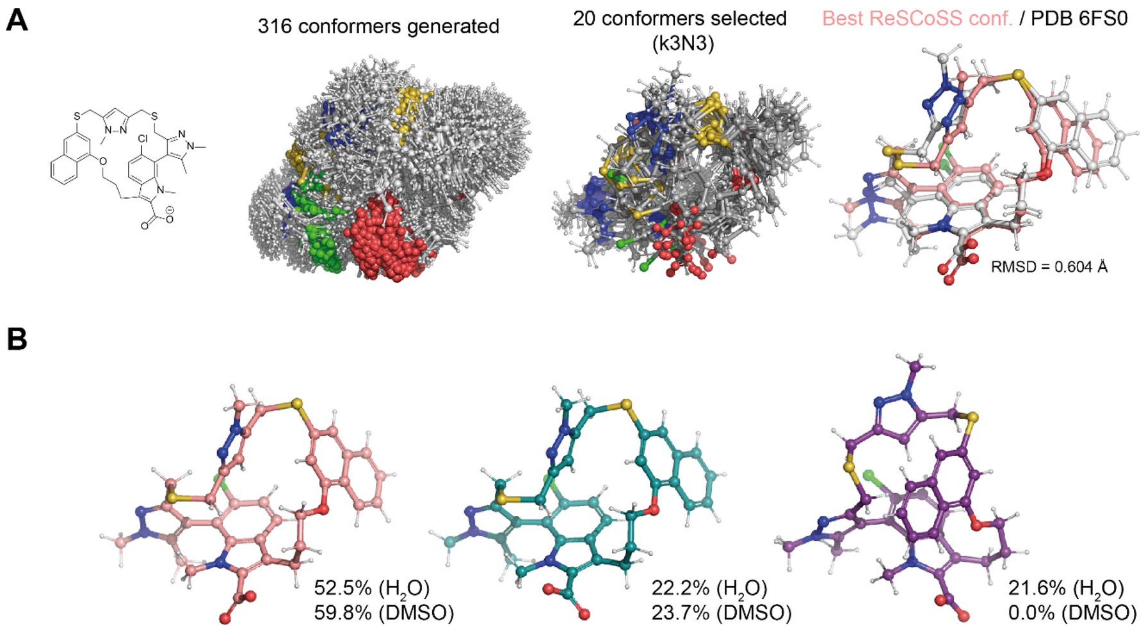


Fig. 6 Application of ReSCoSS to the macrocyle AZD5991. **a** ReSCoSS conformer generation and selection procedure identifies the experimentally known dominant solution conformer out of 316 (generated) and 20 (selected with k3N3) conformers. **b** Top-ranked conformers in water and DMSO with respective Boltzmann weights. The conformation in purple corresponds to the non-bioactive conformation discussed in [50]. The experimentally validated conformer is identified by ReSCoSS as the dominant one in both solvents

in the original manuscript as the global minimum structure is also recapitulated by ReSCoSS as third-most prevalent conformation in water with a weight of 21.6% (ΔΔG = + 0.53 kcal/mol); it is not ranked as relevant in DMSO where it lies + 2.73 kcal/mol above minimum (Fig. 6b). Everything considered, the performance of ReSCoSS on AZD5991 is very convincing both with regards to choosing the relevant conformers for minimization out of a very diverse set and with regards to the free energy assessment of the final set of conformers.

Applying ReSCoSS in property prediction

In our in-house workflows, ReSCoSS is used in a dual capacity: gaining insight into the conformational landscape of a molecule of interest, whether already synthetized or a novel virtual design, and at the same time predicting properties of interest such as logP/D, pKa, or free energy of solvation. These properties are easily obtained using the .cosmo files computed at the COSMO-RS FINE19 level, which are generated at the end of each ReSCoSS run. We recently applied ReSCoSS in our submission to the SAMPL6 blind challenge for the prediction of pKa values where it gratifyingly ranked among the top quantum chemistry based submissions and 4th overall [19]. The prediction of ionization constants depends on quite a number of different factors, among which the selection of relevant conformations is an important factor but by no means the only one. In addition, the level of theory and the choice of linear free energy relationship (LFER) fit have a considerable influence on the resulting pKa prediction. The winning submission in the SAMPL6 challenge, a quantum chemistry based one from the Grimme group, relied on using a double-hybrid DFT gas phase energies combined with explicit modelling of thermochemical contributions and COSMO-RS solvation and a refit of the LFER parameters. This level of theory, while certainly more accurate than COSMO-RS FINE19, requires runtimes which we deem too inefficient in the industry environment, at least given the current state of typical in-house computational capacity.

Assured by the good performance of ReSCoSS for pKa prediction in the SAMPL6 challenge, we turned to two other physicochemical properties that can be predicted using the COSMO-RS framework: The free energy of hydration, ΔG(hyd), and the octanol/water partition coefficient, logP(o/w). logP(o/w) is routinely measured and calculated during lead optimization campaigns as well as pharmaceutical development. Both properties are expected to be at least partially influenced by the selection of conformers and hence serve as an indirect benchmark to assess the conformer selection strategy of ReSCoSS.

Mobley and Guthrie have compiled an excellent free repository of measured hydration free energies for neutral small molecules, FreeSolv [51]. This dataset contains 643 molecules, most of which are quite small, with ΔG(hyd) measured between − 25.5 and 3.4 kcal/mol. Using just a single extended 3D conformation generated by CORINA [26, 27] followed by B97-3c/COSMO optimization and calculation of ΔG(hyd) at the COSMO-RS FINE19 level already results in a very good agreement with experimental values (Fig. 7a) with an MAE of 0.74 kcal/mol. Using ReSCoSS conformers and the k3N3 combination in the selection step further improves the result with the MAE decreased to 0.63 kcal/mol and R^2 = 0.94 (Fig. 7b). Notably, both COSMO-RS based predictions outperform the MD-based

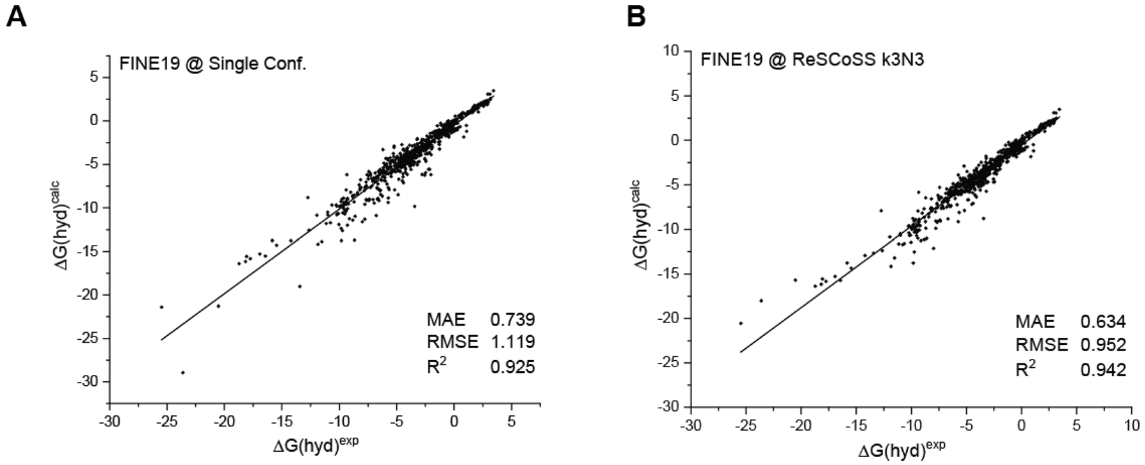


Fig. 7 Application of ReSCoSS to the prediction of free energies of hydration. The 643 experimentally determined ΔG(hyd) values from the FreeSolv v0.51 dataset are plotted against; **a** calculated ΔG(hyd) using COSMO-RS FINE19 based on a single extended conformation optimized with B97-3c/COSMO; **b** calculated ΔG(hyd) using COSMO-RS FINE19 based on ReSCoSS conformers (k3N3) followed by B97-3c/COSMO optimization

approach outlined in the original publication of FreeSolv [51] which shows a MAE of 1.12 kcal/mol and R^2 = 0.87 to experiment.

The octanol/water partition coefficient, logP(o/w) (as well as the related logD that additionally takes ionization into account) is one of the most important parameters used in driving lead optimization and developability assessment. It simultaneously has bearing on aqueous solubility and permeation properties and therefore it is no surprise that together with on-target affinity, logP(o/w) features prominently in drug likeness estimates like LLE [52] and lipE [53]. From a conformer selection standpoint, logP(o/w) is one of the most interesting properties because the relevant solution-phase minimum-energy conformations can often differ between the polar water phase and the rather apolar octanol phase and therefore the ReSCoSS selection strategy is expected to significantly impact the calculated logP(o/w) compared to only using a single conformer per compound. As comprehensive logP(o/w) sets on drug-like molecules measured under the same conditions are rare in literature, we experimentally determined the partition coefficients for 25 approved drugs in this work. This ensures consistent logP(o/w) measurements in the same lab under the same conditions and hence represents an ideal dataset to assess the performance of ReSCoSS in logP(o/w) prediction. We then computed different established fast logP metrics as well as COSMO-RS FINE19 logP(o/w) calculated on single conformers and also using ReSCoSS conformers. The experimental logP range for these drugs spans over 5.5 log units (logP = 0.43 to 6.00) and a comparison of experimental with calculated data is given in Table 2. In addition we computed uncertainties in R^2, RMSD and MAD and report the approximate 95% confidence intervals for each of the parameters as well.

Fast in silico estimators like clogP and RDKit molLogP show decent performance on this dataset with RMSDs around 0.8 log units but both can deviate by over 2 log units in some cases. This is in line with our in-house experience where we use these fast estimators for general application and run ReSCoSS analyses where the standard tools do not show a good performance. It is worth nothing that simply using COSMO-RS to calculate logP(o/w) based on a single conformation does not lead to generally improved predictions over the fast methods. Indeed, with an RMSD of close to 1 unit and R^2 = 0.73, the single-conformer approach is inferior to clogP and molLogP on this dataset. Expanding and selecting out of the tautomer and conformer space with ReSCoSS improves the performance of COSMO-RS FINE19 significantly, with RMSD reduced to 0.64 and R^2 = 0.88 achieved using the k3N3 combination (Fig. 8). In Fig. 9, three concrete examples from the dataset are analyzed. For Mirdametinib (Fig. 9a), the logP is underestimated using the single-conformation approach (calculated 1.18, measured 3.06). Looking at the highest ranked three conformations in both octanol and water phase from the ReSCoSS run indicates why. In octanol, Mirdametinib is predicted to adopt an extended conformation with the polar end of the molecule bearing two hydroxyl groups on one end, and the very hydrophobic iodophenyl moiety at the other. This is in line with the nature of octanol as a hydrophobic elongated molecule with a polar head group. In water, Mirdametinib is predicted to be predominantly folded with the polar head groups shielding the iodophenyl moiety from the polarity of surrounding water. These two classes of conformers are quite distinct and both are needed to reach a good accuracy of the COSMO-RS prediction. The single conformation generated by CORINA, on the other hand, bears similarity to the water conformations but is less compact, while it is distinct from the octanol ones, and so the logP is predicted too low, i.e., the molecule is predicted to be more "happy" in water than it is in reality. Moexipril (Fig. 9b) is another example where ReSCoSS significantly improves the predictions, albeit for a different reason. Moexipril contains a carboxylic acid as well as secondary amine so it can exist in its net neutral form either as a zwitterionic molecule with the acid deprotonated and the amine protonated or as an uncharged neutral molecule. As described previously, ReSCoSS allows for several tautomeric forms of a molecule to be present during the run. Each tautomeric form is conformer-expanded followed by GFN2-xTB/GBSA minimization and B97-3c/COSMO single points. The following cluster step then selects diverse conformations from the combined set for further optimization. Since the tautomers (in this case: zwitterion and neutral form) usually have different properties in terms of dipole moment, hydrogen bond acceptor and donor strength, they fall into different clusters and representatives of each tautomer survive the clustering process if they are reasonably low in energy in any given solvent. In the case of Moexipril, the zwitterionic form clearly dominates in the aqueous phase but the neutral form dominates in the octanol phase, therefore causing any prediction based on only one of the two forms to be considerably wrong. ReSCoSS, taking both forms into account, achieves good predictivity in this case. A third example, Bicalutamide (Fig. 9c), shows the challenges that can hamper COSMO-RS predictions even when the relevant conformers have likely been correctly identified. Bicalutamide has been crystallized both bound to its target, the Androgen receptor (PDB: 1Z95), as well as off-targets such as human serum albumin (PDB: 4LA0) and the human CYP46A1 P450 enzyme (PDB: 4FIA). The bound conformation of bicalutamide to both HSA and P450 is a compact conformation with intramolecular π-stacking. ReSCoSS identifies a perfect overlay to the HSA-bound conformation as the major conformer in water. More extended conformations are present in the final of set of conformers but only one of them, the third-ranked conformation in octanol

Table 2 Calculated logP(o/w) for 25 approved drugs in comparison with experimental data

Drug	logP (exp.)	logP (FINE19, SC)[a]	logP (ReSCoSS k3N3)[b]	logP (ReSCoSS k5N2)[b]	clogP	RDKit MolLogP
Amprenavir	2.42	3.00	2.87	2.87	3.29	2.40
Procyclidine	4.80	5.70	5.70	5.74	4.59	3.94
Semagacestat	1.01	0.52	1.28	1.21	1.66	0.38
Tofacitinib	1.09	0.29	0.53	0.62	1.52	1.06
Dasatinib	3.36	1.95	3.85	3.95	2.53	3.31
Ezetimibe	4.60	4.90	4.31	4.60	3.96	4.89
Cefamandole	0.43	− 0.30	1.42	1.41	0.11	− 0.23
Navarixin	2.80	3.69	3.21	3.23	1.46	2.90
Bicalutamide	2.93	4.15	1.72	1.74	2.71	2.88
Abacavir	1.36	0.82	1.12	1.05	0.81	1.09
Moexipril[c]	0.90	3.79	0.62	1.42	1.39	2.58
Alogliptin	0.65	− 0.12	− 0.36	− 0.35	0.99	0.39
Tioconazole	5.30	4.99	4.63	4.58	4.79	5.86
Begacestat	2.80	3.21	3.21	3.19	2.31	2.78
Regorafenib	6.00	5.70	6.07	6.02	5.19	5.69
Simvastatin	4.50	4.62	3.50	3.50	4.48	4.59
Mirdametinib	3.06	1.18	2.99	2.53	3.00	2.47
Pazopanib	1.60	3.12	2.74	2.75	3.65	3.14
Fluconazole	0.50	0.72	0.61	0.62	− 0.44	0.74
Pevonedistat	2.85	3.15	2.96	2.89	1.16	2.06
Lurasidone	5.80	5.71	5.81	5.84	5.61	4.26
Entinostat	1.19	2.28	1.60	1.61	0.82	3.34
CI-1040	5.20	4.94	5.41	5.39	5.97	5.04
Erlotinib	3.30	3.39	3.64	3.63	4.34	3.41
Roflumilast	3.83	5.09	5.02	5.02	3.00	5.03
MAD		0.775	0.514	0.529	0.666	0.567
MAD, 95% CI		0.544–1.028	0.360–0.660	0.390–0.684	0.495–0.848	ND
RMSD		1.007	0.638	0.653	0.819	0.824
RMSD, 95% CI		0.674–1.334	0.478–0.778	0.495–0.791	0.594–1.055	ND
R^2		0.733	0.878	0.871	0.797	0.782
R^2, 95% CI		0.476–0.893	0.796–0.939	0.768–0.938	0.609–0.913	ND

[a] COSMOtherm logP(o/w) (TZVPD-FINE19) using a single 3D conformation (CORINA) optimized at B97-3c/COSMO level

[b] COSMOtherm logP(o/w) (TZVPD-FINE19) using ReSCoSS conformations selected with k = 3, N = 3/k = 5, N = 2, optimized with B97-3c/COSMO

[c] Moexipril exists as Zwitterion in aqueous phase, logP is given as logD(max) = 0.9

(18.4%), is predicted relevant at the COSMO-RS FINE19 level. This leads to the logP being underestimated by 1.2 units (predicted 1.72, measured 2.93). In contrast, the single conformation from CORINA is extended and leads to a prediction of logP(o/w) = 4.15 which is also off by 1.2 units but in the opposite direction. This relatively large deviation of the COSMO-RS prediction based on ReSCoSS conformers, the largest in the whole set of 25 molecules, may be caused either by the COSMO-RS parameterization, the underlying imprecision of the selected level of DFT theory (BP86-D3(BJ)/TZVPD) or by the fact that thermochemical contributions including entropy are not explicitly modelled. It is of course possible to tackle some of these potential sources of error, e.g. by computing thermochemical contributions explicitly and conducting single point calculations in gas phase at a higher level of theory; we have recently employed this strategy for the prediction of pKa values [19] with success and a similar improvement may be expected for logP. On the other hand, it has also been shown that changing the underlying density functional does not necessarily improve COSMO-RS performance in general [54]. In addition, especially for very polar molecules one could alternatively take one or more solvent molecules explicitly into account during the conformer generation procedure. In conclusion, we note that the COSMO-RS logP(o/w) predictions are clearly and significantly improved by the inclusion

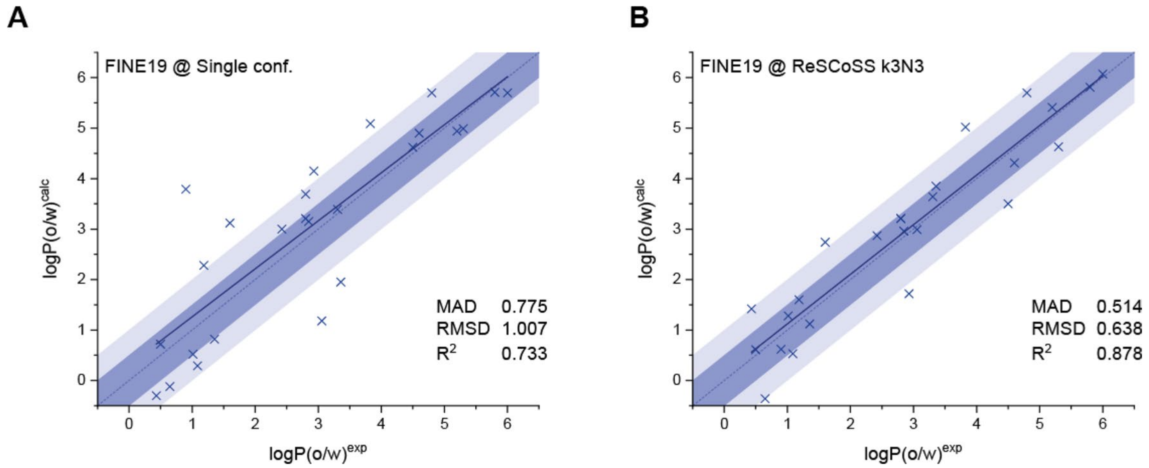


Fig. 8 Application of ReSCoSS to the prediction of logP(o/w). We experimentally determined logP(o/w) for 25 drugs (Table 2). **a** Calculated vs. experimental logP(o/w) using COSMO-RS FINE19 based on a single extended conformation optimized with B97-3c/COSMO; **b** calculated vs. experimental logP(o/w) using COSMO-RS FINE19 based on a set of conformers based on ReSCoSS conformers (k3N3) followed by B97-3c/COSMO optimization. Dark blue and light blue shading indicate 0.5 and 1 log units deviation from identity, respectively

of the ReSCoSS conformations such that only two predictions out of 25 exceed an error margin of 1 log unit.

Conclusions and outlook

We have developed ReSCoSS, a quantum-chemistry based workflow designed to produce small sets of relevant solution conformers for drug-like molecules. ReSCoSS uses the COSMO-RS/COSMOtherm framework to predict conformer weights in solvents of interest defined by the user and can be further used in the prediction of other COSMO-RS properties such as logP(o/w) among many others. Through several case studies and real-life industry applications, we have shown that ReSCoSS is able to correctly identify relevant conformer subsets and that the inclusion of ReSCoSS conformers in COSMO-RS based property predictions has a significant positive effect.

At Novartis, the workflow has found adoption in property prediction for ADME optimization as well as structure-based and ligand-based drug design in early research, but is also used in drug development, allowing a quantum-chemistry perspective on active pharmaceutical ingredients, solid form design and formulation challenges.

Over the past few years, the usage of ReSCoSS as a “computational assay” is emerging: its accuracy allows triaging of in silico ideas, helps avoid syntheses of compounds with suboptimal properties, and in some projects has even been shown to produce prediction accuracy on par with experimental assays. It thus demonstrates the great potential of bringing quantum chemistry methods more prominently to the fore in drug discovery and development. In future publications, we will demonstrate the utility of ReSCoSS in tackling further challenges in the pharmaceutical context, for example the analysis of solvent-mediated conformational polymorphism and use of the workflow for permeability prediction.

Computational details

Starting from 2D structures, 3D conversion was done using the CORINA software [28]. Conformer searches were carried out using Schrodinger Macromodel release 2019.3 [29] with the Monte Carlo multiple minimum (MCMM) method, the OPLS2005 force field including the GBSA implicit solvation model for water. The all-atom RMSD threshold for detections of duplicates was set to 0.75 Å and the potential energy cutoff set at 30 kJ/mol. Additionally, we used CCG MOE 2019.01 [31] to conduct LowModeMD searches [30] with standard settings using the MMFF94x force field, the Born solvation model, the RMSD increased to 0.75 Å with hydrogen detection switched on (all-atom RMSD). All conformers were then combined and duplicate conformers identified and discarded using the GetConformerRMS code in RDKit and a cutoff of 0.5 Å. The remaining conformers were then fully optimized at the GFN2-xTB/GBSA(water) level [34] in the standalone *xtb* code v.6.2.1 [55].

All further QM calculations were carried out using Turbomole v. 7.3 [37, 38]. Single point calculations as well as

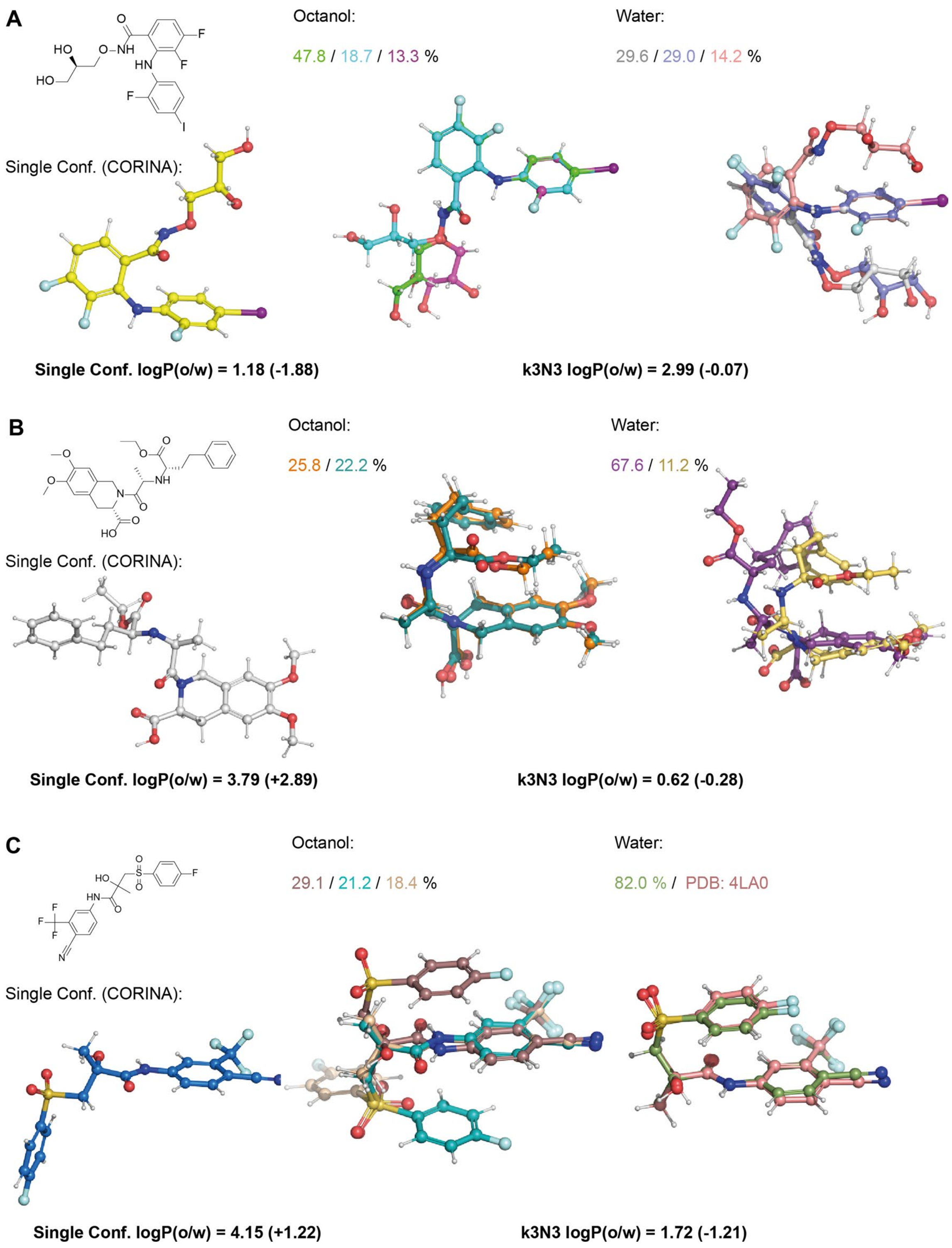


Fig. 9 Comparison of single 3D conformation generated by CORINA with the most relevant conformations in octanol and water for **a** Mirdametinib, **b** Moexipril and **c** Bicalutamide, with associated logP(o/w) values

full geometry optimizations after selection of subsets were carried out using the B97-3c composite method [35] with COSMO solvation [16]. After geometry optimization, a BP86 [56, 57]/def2-TZVPD [58, 59]/COSMO single point calculation including empirical D3(BJ) dispersion correction [60, 61] was carried out for all conformers. The resulting.

cosmo files were used in calculations with COSMOtherm release 19.0.1 [42] at the FINE19 level using standard settings. All molecular depictions were prepared using PyMOL [62]. Violin plots and 3D plots were prepared with Matplotlib [63]. The correlation plots in Figs. 7 and 8 were prepared in Origin 2019 [64].

For the logP(o/w) set we computed uncertainties in R^2, RMSD and MAD implementing a bootstrapping procedure in Python. In this procedure, we considered the pairs (experimental, predicted) as sample of interest and we constructed the bootstrap resample by sampling with replacement from these pairs uniformly at random. We used 1000 replicates and a new linear regression model was fitted to each sampled data, yielding a bootstrap statistic for R^2, RMSD and MAD. We report the approximate 95% confidence intervals for each of the parameters in Table 2. The confidence interval extends from the 2.5th percentile to the 97.5th percentile.

Miniaturized Shake-Flask logP determination

The 1-octanol/water partitioning coefficient (logP) was determined using a miniaturized Shake-Flask equilibrium method adapted from [65]. logD was measured in three different buffers of varying pH (2, 7.4 and 11) and the logP was extracted from the measured logD at the pH where the compound is neutral. Prior to starting the experiment the two phases were pre-saturated, so "water-saturated 1-octanol" and "1-octanol-saturated water" were used. The samples were initially dissolved in DMSO as a 10 mM stock concentration. The samples and an internal standard were dispensed in a 1 ml deepwell plate and DMSO was evaporated prior to dissolution in 1-octanol at a target concentration of 150 µM while shaking at 1000 rpm during 8 h. The buffers were added with a phase ratio K of 1 (where K = Vwater/Voctanol) and then the samples were shaken 4 h on a shaker at 1000 rpm. The deepwell plate was then centrifuged at 3000 rpm prior to phase separation. A × 10 dilution for the aqueous phase and a × 1000 dilution for the octanol phase were prepared and quantified by LC-HRMS against an internal standard (Dexamethasone) with a known logD = 1.9 using the following equation:

$$\log D = \log\left(\frac{Analyte\ peak\ area\ in\ octanol * 1000/IS\ peak\ area\ in\ octanoll/0.794}{Analyte\ peak\ area\ in\ aqueous * 10/IS\ peak\ area\ in\ aqueous}\right)$$

Column used: Zorbax_SB_AQ 50 × 2.1 mm 1.8 µm – Column oven temperature = 50 °C.

Mobile phase: A = 100% water UHPLC grade + 0.08% Formic acid. B = 100% ACN + 0.08% Formic acid. Flow rate = 0.5 ml/min. Gradient mode: starting at 95% A up to 95% B in 0.5 min and kept constant during 1 min before to restore initial conditions within 0.1 min. Vinj = 5 µl. Full positive acquisition mode—Full scan 130 to 1800 m/z and Resolution = 35,000.$[M + H]^+$ ion chromatogram was extracted for each compound. This protocol was followed for all compounds except Navarixin and Moexipril.

Potentiometric logP determination

The partitioning coefficients for Navarixin and Moexipril were determined on the commercial SiriusT3 instrument (Pion-inc.com) as described by Avdeef [66]. Briefly, 0.5 to 1 mM of test solutions were titrated from pH 2 to 12 for bases or 12 to 2 for acids. Titrations were conducted at 25 °C and in 0.15 M ionic strength. Aqueous titrations were performed in triplicate in 0.15 M KCl. A minimum of three titrations in varying amounts of octanol as partitioning solvent were performed for extracting the logP information. For each titration, initial estimates of apparent pKa values were obtained from Bjerrum difference plots (number of bound protons versus pH) and then were refined by the instrument software.

Acknowledgements We thank colleagues from the global CADD group at NIBR-GDC and Grahame Woollam (TRD) for valuable feedback on applying ReSCoSS in projects. Wolfgang Zipfel and the NIBR NX Scientific Computing team are thanked for generous allocation of HPC resources. We thank Stefan Grimme for discussions around xTB, DFT functionals and solvation, Uwe Huniar for advice on Turbomole and Jens Reinisch and Michael Diedenhofen for helpful discussions about conformations and COSMO-RS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants and/or animals No animal or human clinical studies were performed as part of this work.

References

1. Kuhn B, Guba W, Hert J, Banner D, Bissantz C, Ceccarelli S, Haap W, Körner M, Kuglstatter A, Lerner C, Mattei P, Neidhart W, Pinard E, Rudolph MG, Schulz-Gasch T, Woltering T, Stahl M (2016) J Med Chem 59(9):4087
2. Persch E, Dumele O, Diederich F (2015) Angew Chem Int Ed 54(11):3290
3. Vippagunta SR, Brittain HG, Grant DJW (2001) Adv Drug Deliv Rev 48(1):3
4. Cruz-Cabeza AJ, Bernstein J (2014) Chem Rev 114(4):2170
5. Abramov YA, Zhang P, Zeng Q, Yang M, Liu Y, Sekharan S (2020) Cryst Growth Des 20(3):1512
6. Boström J (2001) J Comput Aided Mol Des 15(12):1137
7. Hawkins PCD (2017) J Chem Inf Model 57(8):1747

8. Friedrich N-O, Meyder A, de Bruyn KC, Sommer K, Flachsenberg F, Rarey M, Kirchmair J (2017) J Chem Inf Model 57(3):529
9. Friedrich N-O, de Bruyn KC, Flachsenberg F, Sommer K, Rarey M, Kirchmair J (2017) J Chem Inf Model 57(11):2719
10. Gürsoy O, Smieško M (2017) J Cheminform 9(1):29
11. Cavasin AT, Hillisch A, Uellendahl F, Schneckener S, Göller AH (2018) J Chem Inf Model 58(5):1005
12. Kanal IY, Keith JA, Hutchison GR (2018) Int J Quant Chem 118(5):e25512
13. Folmsbee D, Hutchison G (2020) ChemRxiv. https://doi.org/10.26434/chemrxiv.11920914.v2
14. Klamt A (1995) J Phys Chem 99(7):2224
15. Klamt A, Jonas V, Bürger T, Lohrenz JCW (1998) J Phys Chem A 102(26):5074
16. Klamt A (2018) WIREs Comput Mol Sci 8(1):e1338
17. Sure R, Grimme S (2015) J Chem Theory Comput 11(8):3785
18. Klamt A, Eckert F, Reinisch J, Wichmann K (2016) J Comput Aided Mol Des 30(11):959
19. Pracht P, Wilcken R, Udvarhelyi A, Rodde S, Grimme S (2018) J Comput Aided Mol Des 32(10):1139
20. Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) J Phys Chem A 107(44):9380
21. Loschen C, Reinisch J, Klamt A (2020) J Comput Aided Mol Des 34(4):385
22. Perola E, Charifson PS (2004) J Med Chem 47(10):2499
23. Günther S, Senger C, Michalsky E, Goede A, Preissner R (2006) BMC Bioinform 7(1):293
24. Klamt A, Eckert F, Diedenhofen M (2009) J Phys Chem B 113(14):4508
25. COSMOconf 4.0, COSMOlogic GmbH & Co KG, a Dassault Systèmes company
26. Sadowski J, Gasteiger J, Klebe G (1994) J Chem Inf Comput Sci 34(4):1000
27. Schwab CH (2010) Drug Discov Today 7(4):e245
28. Molecular Networks GmbH: 3D structure generator CORINA classic. Nuremberg, Germany. www.mn-am.com. Accessed 4 June 2020
29. Schrödinger Release 2019-3: MacroModel, Schrödinger, LLC, New York (2020)
30. Labute P (2010) J Chem Inf Model 50(5):792
31. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, Montreal, QC, Canada
32. Riniker S, Landrum GA (2015) J Chem Inf Model 55(12):2562
33. RDKit: open-source cheminformatics. https://www.rdkit.org. Accessed 4 June 2020
34. Bannwarth C, Ehlert S, Grimme S (2019) J Chem Theory Comput 15(3):1652
35. Brandenburg JG, Bannwarth C, Hansen A, Grimme S (2018) J Chem Phys 148(6):064104
36. Klamt A, Schüürmann G (1993) J Chem Soc Perkin Trans 2(5):799
37. TURBOMOLE 7.3, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007. https://www.turbomole.com. Accessed 4 June 2020
38. Furche F, Ahlrichs R, Hättig C, Klopper W, Sierka M, Weigend F (2014) WIREs Comput Mol Sci 4(2):91
39. Xue S-J, Duan L-P, Ke S-Y, Zhu J-M (2004) CCDC 234134: experimental crystal structure determination. Cambridge Crystallographic Data Centre, Cambridge
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) J Mach Learn Res 12(85):2825
41. Eckert F, Klamt A (2002) AIChE J 48(2):369
42. COSMOtherm release 19.0.1, COSMOlogic GmbH & Co KG, a Dassault Systèmes Company
43. Lu J, Wang X-J, Yang X, Ching C-B (2007) Cryst Growth Des 7:9
44. Groom CR, Allen FH (2014) Angew Chem Int Ed 53(3):662
45. Brameld KA, Kuhn B, Reuter DC, Stahl M (2008) J Chem Inf Model 48(1):1
46. Maity S, Jana P, Maity SK, Kumar P, Haldar D (2012) Cryst Growth Des 12(1):422
47. Coleman PJ, Schreier JD, Cox CD, Breslin MJ, Whitman DB, Bogusky MJ, McGaughey GB, Bednar RA, Lemaire W, Doran SM, Fox SV, Garson SL, Gotter AL, Harrell CM, Reiss DR, Cabalu TD, Cui D, Prueksaritanont T, Stevens J, Tannenbaum PL, Ball RG, Stellabott J, Young SD, Hartman GD, Winrow CJ, Renger JJ (2012) ChemMedChem 7(3):415
48. Tamanini E, Buck IM, Chessari G, Chiarparin E, Day JEH, Frederickson M, Griffiths-Jones CM, Hearn K, Heightman TD, Iqbal A, Johnson CN, Lewis EJ, Martins V, Peakman T, Reader M, Rich SJ, Ward GA, Williams PA, Wilsher NE (2017) J Med Chem 60(11):4611
49. Tron AE, Belmonte MA, Adam A, Aquila BM, Boise LH, Chiarparin E, Cidado J, Embrey KJ, Gangl E, Gibbons FD, Gregory GP, Hargreaves D, Hendricks JA, Johannes JW, Johnstone RW, Kazmirski SL, Kettle JG, Lamb ML, Matulis SM, Nooka AK, Packer MJ, Peng B, Rawlins PB, Robbins DW, Schuller AG, Su N, Yang W, Ye Q, Zheng X, Secrist JP, Clark EA, Wilson DM, Fawell SE, Hird AW (2018) Nat Commun 9(1):1
50. Balazs AYS, Carbajo RJ, Davies NL, Dong Y, Hird AW, Johannes JW, Lamb ML, McCoull W, Raubo P, Robb GR, Packer MJ, Chiarparin E (2019) J Med Chem 62(21):9418
51. Mobley DL, Guthrie JP (2014) J Comput Aided Mol Des 28(7):711
52. Leeson PD, Springthorpe B (2007) Nat Rev Drug Discov 6(11):881
53. Ryckmans T, Edwards MP, Horne VA, Correia AM, Owen DR, Thompson LR, Tran I, Tutt MF, Young T (2009) Bioorg Med Chem Lett 19(15):4406
54. Reinisch J, Diedenhofen M, Wilcken R, Udvarhelyi A, Glöß A (2019) J Chem Inf Model 59(11):4806
55. xtb release v6.2.1, 2019. https://github.com/grimme-lab/xtb/. Accessed 4 June 2020
56. Becke AD (1988) Phys Rev A 38(6):3098
57. Perdew JP (1986) Phys Rev B 33(12):8822
58. Weigend F, Ahlrichs R (2005) Phys Chem Chem Phys 7(18):3297
59. Rappoport D, Furche F (2010) J Chem Phys 133(13):134105
60. Grimme S, Antony J, Ehrlich S, Krieg H (2010) J Chem Phys 132(15):154104
61. Grimme S, Ehrlich S, Goerigk L (2011) J Comput Chem 32(7):1456
62. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
63. Hunter JD (2007) Comput Sci Eng 9(3):90
64. Origin(Pro), Version 2019. OriginLab Corporation, Northampton, MA, USA
65. Low YW, Blasco F, Vachaspati P (2016) Eur J Pharm Sci 92:110
66. Avdeef A (1992) Quant Struct-Act Relat 11(4):510

Affiliations

Anikó Udvarhelyi[1] · Stephane Rodde[2] · Rainer Wilcken[2]

✉ Anikó Udvarhelyi
aniko.udvarhelyi@novartis.com

✉ Rainer Wilcken
rainer.wilcken@novartis.com

[1] Technical Research and Development, Novartis Pharma AG, 4002 Basel, Switzerland

[2] Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 4002 Basel, Switzerland