

Towards a SMART workflow for Computational Spectroscopy

Daniele Licari,^{ab#} Marco Fusé,^{a#} Andrea Salvadori,^a Nicola Tasinato,^a Marco Mendolicchio,^a
Giordano Mancini^a and Vincenzo Barone^{a*}

^aScuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy

^bIstituto Italiano di Tecnologia, via Morego 30, 16163 Genova, Italy

[#]These authors contributed equally to the work

Electronic Supporting Information

S1 Examples of advanced hybrid methods

Coordinate	MSR results						Reference
	B2PLYP/VTZ	B2PLYP/VTZ + LRA (bonds)	B2PLYP/VTZ + LRA (bonds and angles)	B3LYP/SNSD	B3LYP/SNSD + LRA (bonds)	B3LYP/SNSD + LRA (bonds and angles)	bestCC
$r(\text{C2} - \text{O2})$	1.350(3)	1.349(2)	1.349(2)	1.347(3)	1.348(3)	1.348(3)	1.3476
$r(\text{C2} - \text{O1})$	1.203(2)	1.203(2)	1.203(2)	1.203(3)	1.204(2)	1.204(2)	1.2021
$r(\text{O3} - \text{H5})$	0.968(7)	0.964(3)	0.964(3)	0.97(2)	0.963(6)	0.963(6)	0.9645
$r(\text{C1} - \text{C2})$	1.513(1)	1.513(1)	1.513(1)	1.513(1)	1.513(1)	1.513(1)	1.5128
$r(\text{C1} - \text{N})$	1.442(2)	1.443(2)	1.443(2)	1.442(3)	1.442(2)	1.442(2)	1.4430
$r(\text{C1} - \text{H3})$	1.0907(3)	1.0907(3)	1.0907(3)	1.0907(3)	1.0907(3)	1.0907(3)	1.0903
$r(\text{N} - \text{H1})$	1.011(7)	1.010(3)	1.010(3)	1.02(1)	1.011(6)	1.011(6)	1.0109
$a(\text{O1} - \text{C2} - \text{O2})$	123.1(1)	123.0(1)	123.0(1)	122.3(2)	123.0(1)	123.0(1)	123.05
$a(\text{C1} - \text{C2} - \text{O2})$	111.5(2)	111.5(2)	111.5(2)	111.5(2)	111.5(2)	111.5(2)	111.35
$a(\text{N} - \text{C1} - \text{C2})$	115.3(1)	115.3(1)	115.3(1)	115.3(2)	115.3(1)	115.3(1)	115.25
$a(\text{H3} - \text{C1} - \text{C2})$	107.38(5)	107.38(5)	107.38(5)	107.37(5)	107.37(5)	107.37(5)	107.49
$a(\text{H1} - \text{N} - \text{C1})$	109.7(6)	109.9(6)	109.9(6)	109.9(6)	110.2(8)	110.0(7)	109.86
$a(\text{H5} - \text{O2} - \text{C2})$	106.6(7)	106.6(8)	106.7(8)	107.1(8)	107.1(7)	106.6(8)	106.64
$d(\text{H4} - \text{C1} - \text{C2} - \text{O1})$	123.21(4)	123.20(4)	123.20(4)	123.21(4)	123.22(4)	123.22(4)	123.18
$d(\text{H2} - \text{N} - \text{C1} - \text{C2})$	57.2(6)	57.4(5)	57.4(5)	57(1)	57.4(6)	57.4(6)	57.93

Table S1. Equilibrium geometry of the Ip Conformer of Glycine (for more details see Ref. 136).

S2 RD-VPT2 approach in a series of Nickel dicarbonyl complexes

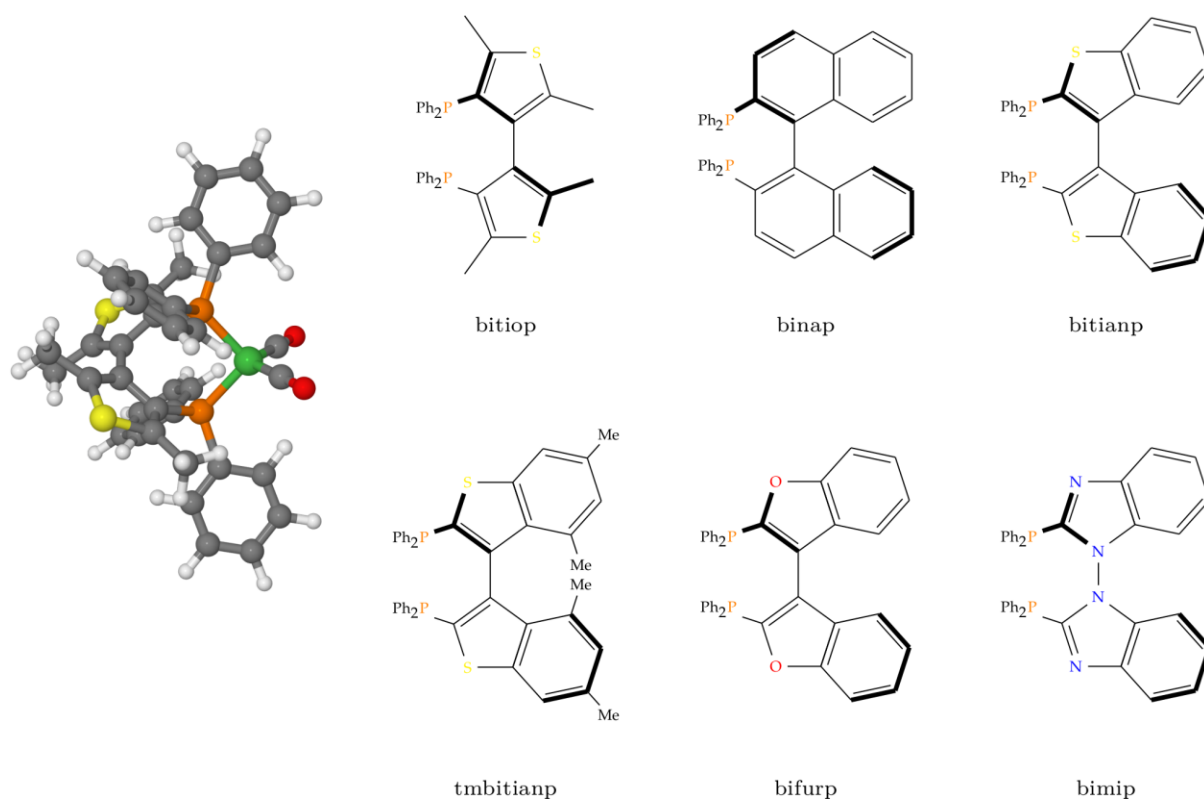


Figure S1 Chemical structure and names of the diphosphines used in reference ¹.

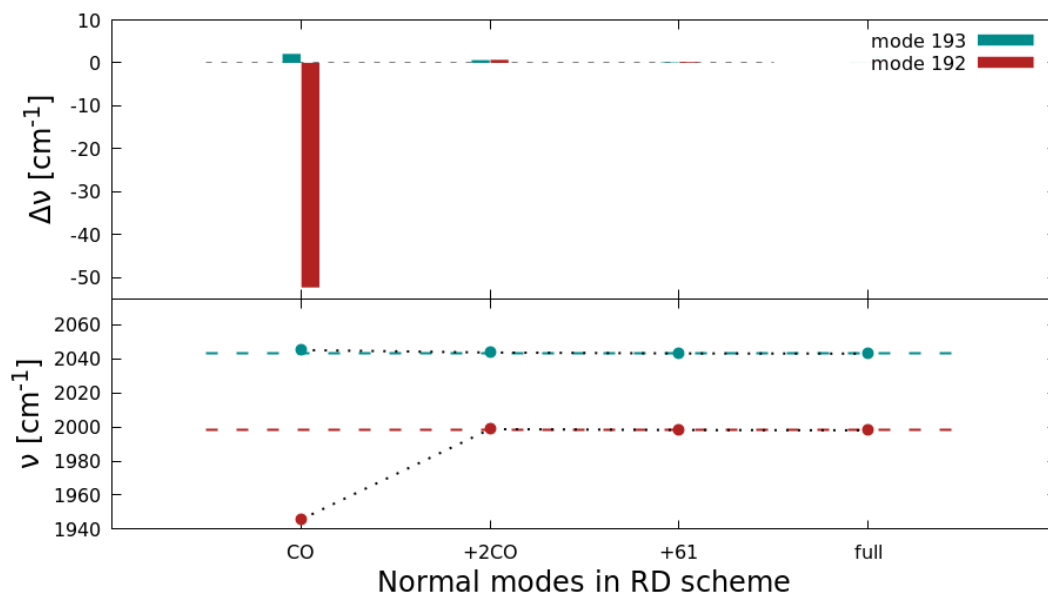


Figure S2 Values of the predicted energies for the two CO stretching (mode 192 and 193) of the $[\text{Ni}(\text{CO})_2(\text{bitiop})]$ complex against the modes included in the RD scheme. The calculated ν_{CO} values refer to gas phase calculations at the B3LYP-D3/6-31+G* level of theory. In the top panel the difference from the full calculation, while in the lower one the value of ν_{CO} are reported. Since the symmetric CO stretching (mode 193) do not show any strong coupling with other modes (see figure 12 in text), it almost converged as only mode included in the RD scheme. On the other

hand, the asymmetric CO stretching (mode 192) required the inclusion of mode 193 to approach the value of the full dimension calculation.

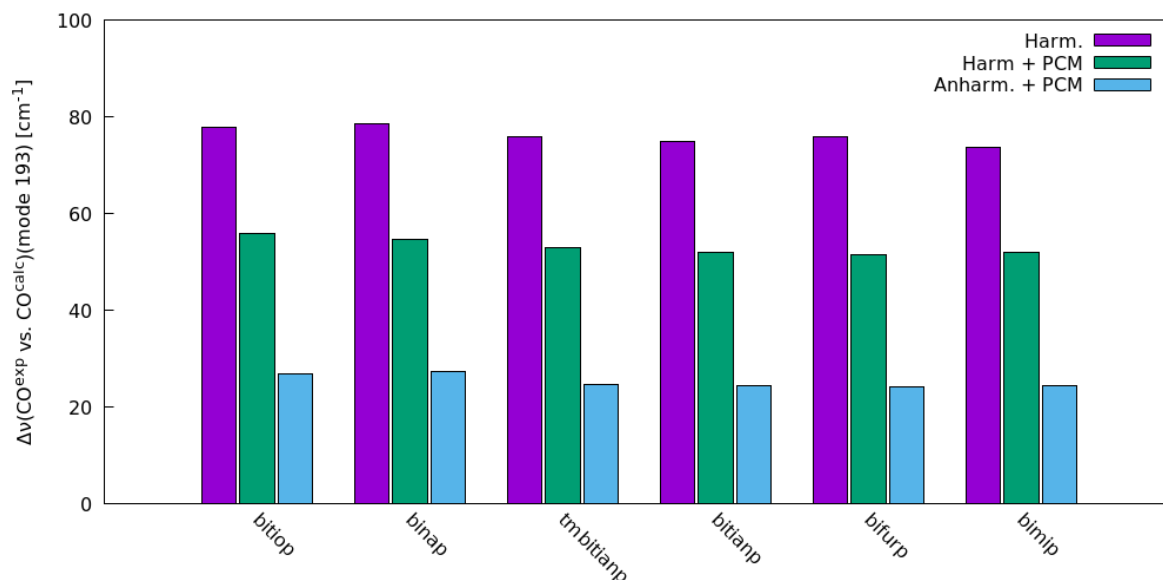


Figure S3 Difference of the calculated ν_{CO} (mode 193) from the experimental value: harmonic calculated carbonyl stretching frequencies in gas phase (violet); including bulk solvent effects (PCM) (green); including both PCM and the anharmonic correction (cyan). All the calculations have been performed at B3LYP-D3/6-31+G* level of theory and the anharmonic calculation refers to RD-VPT2 calculations, which included the two CO stretching modes. Values taken from Bloino et al. (ref. 1)

S3 Brief introduction to vector field visualization

As said in Section 4.3, the simplest method for graphically depicting vector fields consist is the so-called “hedgehog” or “direct” representation, in which vectors are depicted as glyphs, such as lines or arrows. However, a plain hedgehog representation in which every vector of the data set is explicitly depicted is unfeasible in all but the smallest data sets. Therefore, some sort of “simplification” is required, either by adopting a visual representation with an higher level of abstraction or by deriving a more concise and/or informative dataset from the original one. In both cases the goal is to obtain clear and informative images, in which all the meaningful features of the original dataset are preserved and presented. In the following, the others commonly used categories of visual representations for vector fields will be briefly described. Then, a couple of approaches for deriving a more concise and/or informative dataset from the original one will be briefly presented. Apart from the direct glyph-based approach, another widespread class of visual representations for visualizing flow fields is the one making use of the so-called “stream objects”,² such as streamlines, pathlines, streaklines, stream surfaces etc. Stream objects are generated on the base of the path covered by a set of massless particles when inserted in the (steady or unsteady) vector field. The main difficulty when adopting these representations is the choice of the number and of the initial position of the seeding particles: if not properly placed, the resulting stream objects may fail to depict important features; on the other hand, filling the volume with a large number of particles may

result in cluttered images difficult to comprehend. A survey of this class of techniques is presented in McLoughlin et al.³

In Texture-based representations, instead, a texture is generated providing a dense and almost continuous representation of the vector field. Their main advantage is the ability to depict the flow field with a considerable detail and coverage, allowing to easily identify most of the significant features of the field. However, these techniques are particularly suited for visualizing 2D flows or flows over a 3D surface, while they present several perceptual challenges when applied to 3D fields via direct volume rendering. Although methods have been presented to enhance perception of depth, flow direction and relevant features in this scenario (e.g. see Interrante and Gorsh⁴), glyph and stream based techniques “generalize better to 3D fields”.⁵ The simpler possible approach to reduce the size of data set consists in dividing it in regions containing elements having a strong similarity between them and a weak similarity with the elements belonging to the other regions, i. e. by performing a cluster analysis⁶ on the elementary vectors. Once a simplified vector field is derived by means of a clustering procedures, it can then be visualized using a glyphs or stream objects. Over the years many clustering methods has been proposed⁶ and, by being usually agnostic with respect to the nature of the data being clustered (i. e. with which features the similarity / dissimilarity score is calculated and, often, what type of metric is used on this feature space), most of them can be employed for clustering vector fields.

More sophisticated approaches for simplifying vector fields consist in detecting relevant features within the dataset, such as vortices, attracting and repelling points and boundaries between regions of interest. Once detected, these features can be explicitly visualized by means of dedicated glyphs, colored regions and so on. Also vector glyphs and stream objects representations can benefit from features detection: by making an heavier use of graphical elements in proximity of the features and a sparse one in uninteresting regions it is possible to highlight all the relevant characteristics of the field while minimizing cluttering. The reader can refer to Post et al.⁷ and Laramée et al.⁸ for a survey on the state of the art about feature-based flow visualization.

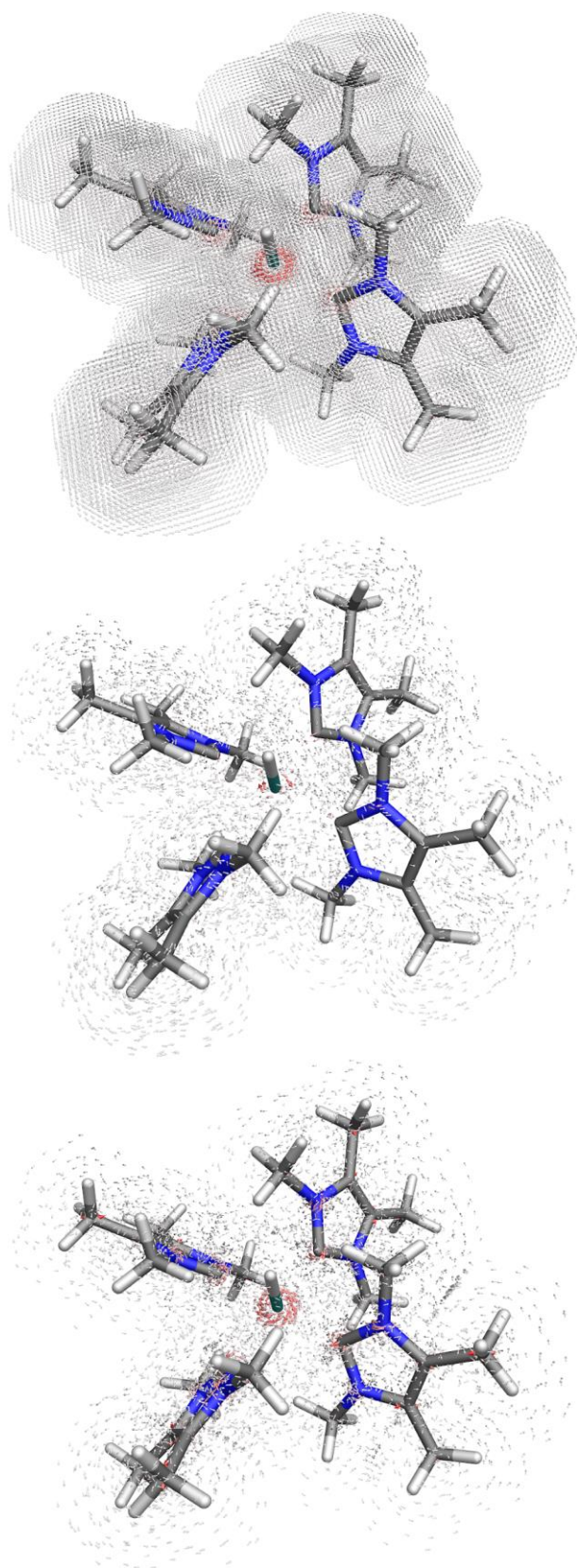


Figure S4 Representations of the same VF dataset by using different simplification algorithms. (Upper panel) One glyph every ten is displayed. Note the confusing “stratification” resulting from the regular sampling. (middle panel) 32000 randomly sampled vectors are displayed. Although the resulting image is much clearer than (a), the magnetically induced current density around Ru-H bond are much less evident with respect to panel c, due to the random subsampling. (lower panel) 16000 representative glyphs obtained with the agglomerative hierarchical clustering algorithm (see text) are

displayed. Note how simplification via clustering provides clear images that preserve the important features of the dataset. In all the cases the same threshold on the vector magnitude have been used.

S4 Uridine Molecular Dynamics simulations and QM calculations

Uridine (see Figure S5) topology was built using the latest version of the ff14SB Amber force field^{9,10} and the put at center of a cubic box made up of 9777 TIP3P molecules. Two short (1 ns) NVT simulations were performed at 150.0 and 298.15 K for thermalization purposes, using an integration time step of 0.1 fs. Then, 50 ns classical NVT MD GROMACS¹¹ simulation was used for the analysis (5000 frames, a 2 fs time step with a velocity-scale thermostat). Bonds were constrained using the LINCS⁴ algorithm. The particle mesh Ewald¹² method was used in the same conditions as in the Tyrosine case as well as the use of the periodic boundary conditions. Vertical transition energies were computed at the CAM-B3LYP¹³/SNSD¹⁴ level of theory on selected configurations sampled during the MD simulations. The obtained energies were convoluted with Gaussian functions in the energy domain using a properly chosen half width at half maximum (HWHM) value, Δ_v :¹⁵

$$\varepsilon_v \propto \sum_{i \in \text{states}} \frac{f_i}{\Delta_v} e^{-\left(\frac{v-v_i^0}{\sigma_v}\right)^2}$$

where

$$\sigma_v = [2\sqrt{2\ln(2)}]^{-1}\Delta_v$$

and f_i and v_i^0 are the oscillator strength and the frequency (in wavenumbers) of the i -th excitation, respectively. Environmental effects were included using C-PCM.¹⁶ The complete TD-DFT data set is shown in Figure S6.

Before performing clustering calculations we have evaluated the performance of USR with respect to RMSD in similar conditions and compared the clusters created using the two dissimilarity measures. Comparing the time needed to obtain the complete 5000x5000 distance matrix, USR was four to five times faster than fitting/RMSD using a simple C implementation of the former with respect to the tools provided by common MD analysis suites (e. g. the GROMACS tools or MDtraj, <http://mdtraj.org/1.9.0/>). Note that the computational cost connected to distance evaluation is negligible when using a several thousand frames but becomes relevant if using a few tenths.

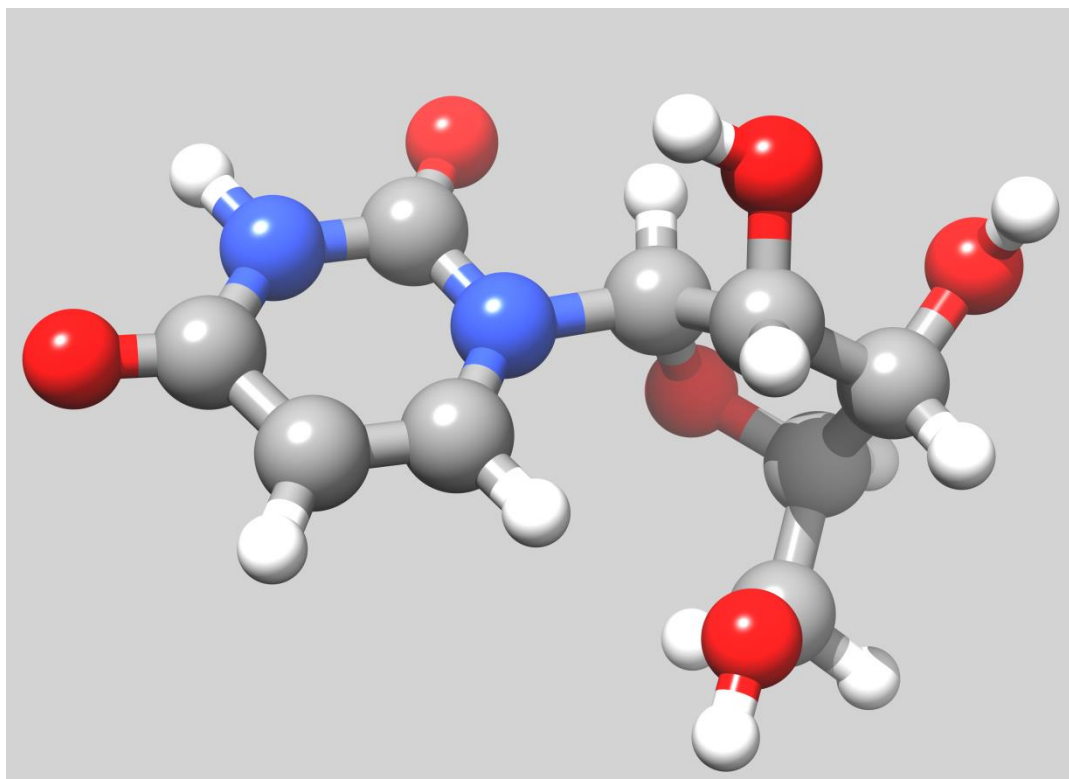


Figure S5. Ball and stick representation of the Uridine molecule.

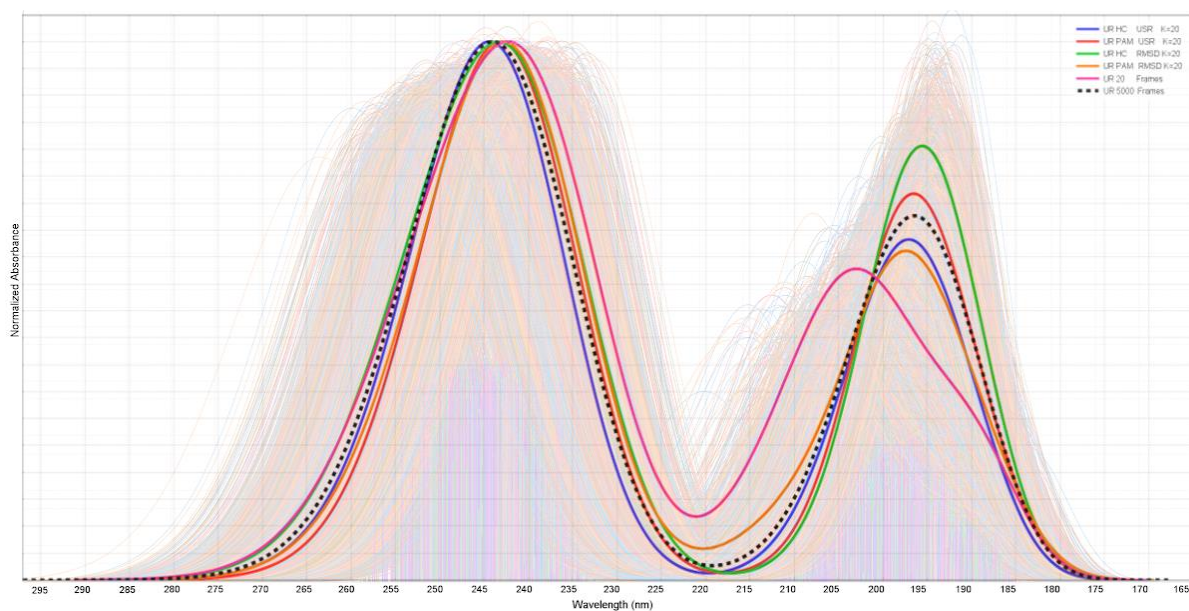


Figure S6. Comparison between final theoretical absorption spectra of uridine in water calculated on 5k frames (dotted blackline), the outcome of clusterings (continuous blue, red, green and orange lines) and 20 randomly selected configurations (continuous pinkline) In the background (in transparency) the absorption spectra of the 5K frames.

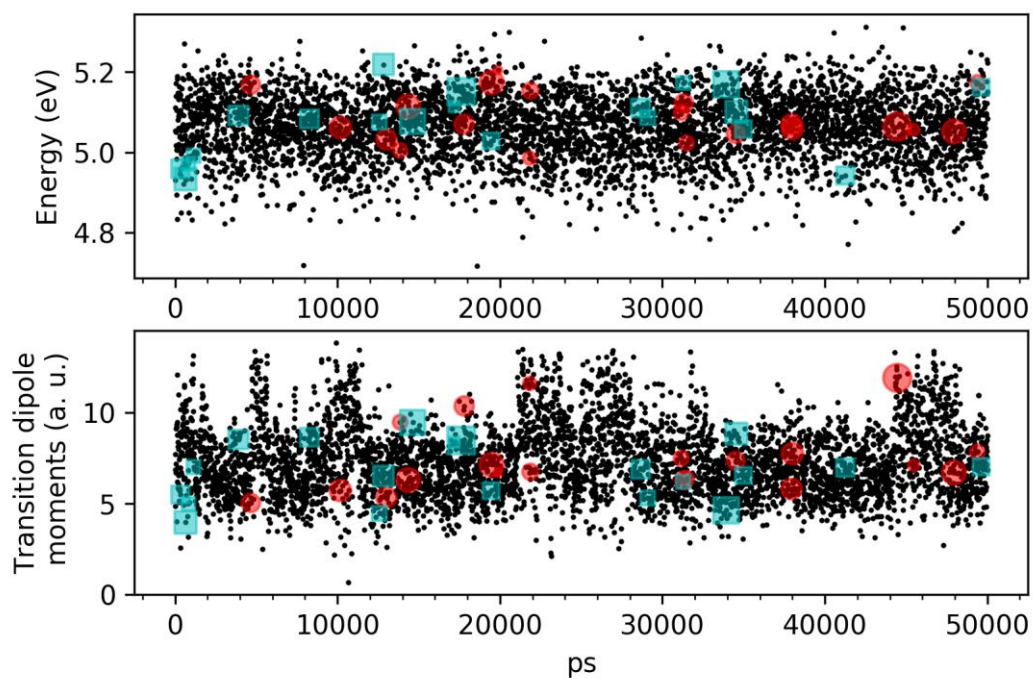


Figure S7. Energies (upper panel) and transition dipole moment modulus (lowe panel) for the first excited state obtained for the complete reference set (black dots), PAM clustering with USR (red circles) and with RMSD (cyan squares). Markers size of centroids is proportional to cluster size.

References

- 1 M. Biczysko, P. Panek, G. Scalmani, J. Bloino and V. Barone, Harmonic and Anharmonic Vibrational Frequency Calculations with the Double-Hybrid B2PLYP Method: Analytic Second Derivatives and Benchmark Studies, *J. Chem. Theory Comput.*, 2010, **6**, 2115–2125.
- 2 A. Telea, *Data visualization: principles and practice*, CRC Press, Taylor & Francis Group, Boca Raton, Second edition., 2015.
- 3 T. McLoughlin, R. S. Laramée, R. Peikert, F. H. Post and M. Chen, Over Two Decades of Integration-Based, Geometric Flow Visualization, *Comput. Graph. Forum*, 2010, **29**, 1807–1829.
- 4 V. Interrante and C. Grosch, Visualizing 3D flow, *IEEE Comput. Graph. Appl.*, 1998, **18**, 49–53.
- 5 R. S. Laramée, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post and D. Weiskopf, The State of the Art in Flow Visualization: Dense and Texture-Based Techniques, *Comput. Graph. Forum*, 2004, **23**, 203–221.
- 6 C. C. Aggarwal and C. K. Reddy, Eds., *Data clustering: algorithms and applications*, Chapman and Hall/CRC, Boca Raton, 2014.
- 7 F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramée and H. Doleisch, The State of the Art in Flow Visualisation: Feature Extraction and Tracking: The State of the Art in Flow Visualisation: Feature Extraction and Tracking, *Comput. Graph. Forum*, 2003, **22**, 775–792.
- 8 R. S. Laramée, H. Hauser, L. Zhao and F. H. Post, in *Topology-based Methods in Visualization*, eds. H. Hauser, H. Hagen and H. Theisel, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 1–19.
- 9 R. Salomon-Ferrer, D. A. Case and R. C. Walker, An overview of the Amber biomolecular simulation package: Amber biomolecular simulation package, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**, 198–210.
- 10 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
- 11 S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, *Bioinformatics*, 2013, **29**, 845–854.
- 12 T. Darden, L. Perera, L. Li and L. Pedersen, New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations, *Structure*, 1999, **7**, R55–R60.
- 13 T. Yanai, D. P. Tew and N. C. Handy, A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP), *Chem. Phys. Lett.*, 2004, **393**, 51–57.
- 14 E. Stendardo, A. Pedone, P. Cimino, M. Cristina Menziani, O. Crescenzi and V. Barone, Extension of the AMBER force-field for the study of large nitroxides in condensed phases: an ab initio parameterization, *Phys. Chem. Chem. Phys.*, 2010, **12**, 11697.
- 15 X.-H. Zhang, L.-Y. Wang, G.-H. Zhai, Z.-Y. Wen and Z.-X. Zhang, The absorption, emission spectra as well as ground and excited states calculations of some dimethine cyanine dyes, *J. Mol. Struct. THEOCHEM*, 2009, **906**, 50–55.
- 16 M. Cossi, N. Rega, G. Scalmani and V. Barone, Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model, *J. Comput. Chem.*, 2003, **24**, 669–681.