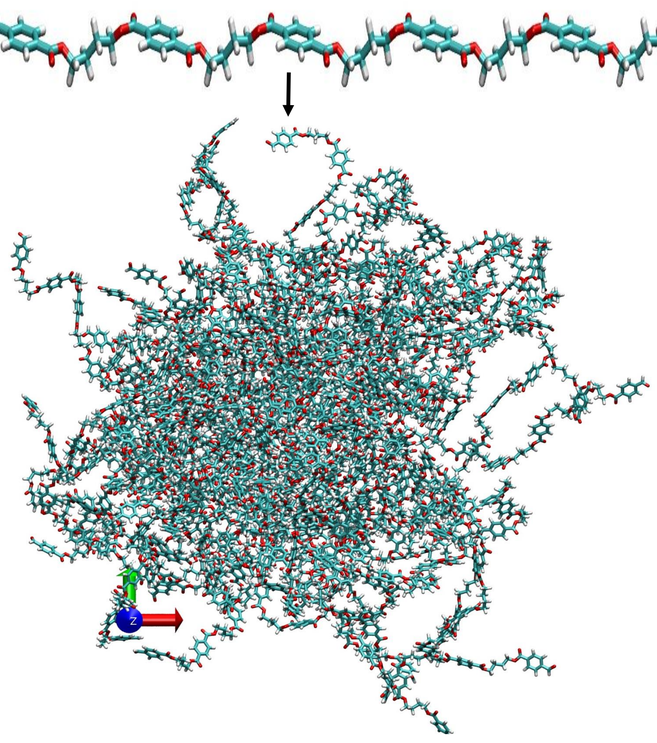
High throughput screening of polymers for properties prediction

Dr Vittoria Fantauzzo, Prof Alessandro Troisi
University of Liverpool

[*]OCCCCOC(=O)c1ccc(C(=O)[*])



Derive a workflow that produces **robustly** and **reproducibly equilibrated** models of polymers starting from the SMILES representation of monomer

1- From SMILEs to simulation

2- Equilibration test

3- Analysis

1- From SMILEs to simulation

monomer SMILEs
- Initial dataset contains 100 polymers[1]

oligomers' structures
- *PSP -> Polymer Structure Prediction*[2] Software to create the oligomer

oligomers' FF
- *LigParGen*[3] Force-field: OPLS-AA
- Input files for GROMACS

simulation box
- Max chain length: 200 atoms
- Box: 15,000 total atoms

simulation
- Geometry optimisation
- MD

2 -Equilibration test: ΔRDF

Variation of the intermolecular C-C RDF between consecutive annealing cycles ($r_{max} = 2$)
If there is no significant variation, the system has reached convergence

$$\Delta RDF_n = \frac{1}{r_{max}} \int_0^{r_{max}} RDF_{(n+1)}(r) - RDF_n(r)\, dr$$

Figure 1. Variation of the distribution of ΔRDF 1-2 and 2-3. Includes example of three cases of ΔRDF between cycle one, two and three of annealing. Left and middle picture show negligible and little variation, while right picture shows extreme variance, with RDF of cycle 1 of annealing being evidence of an error in the simulation. Threshold arbitrarily set to 0.03.

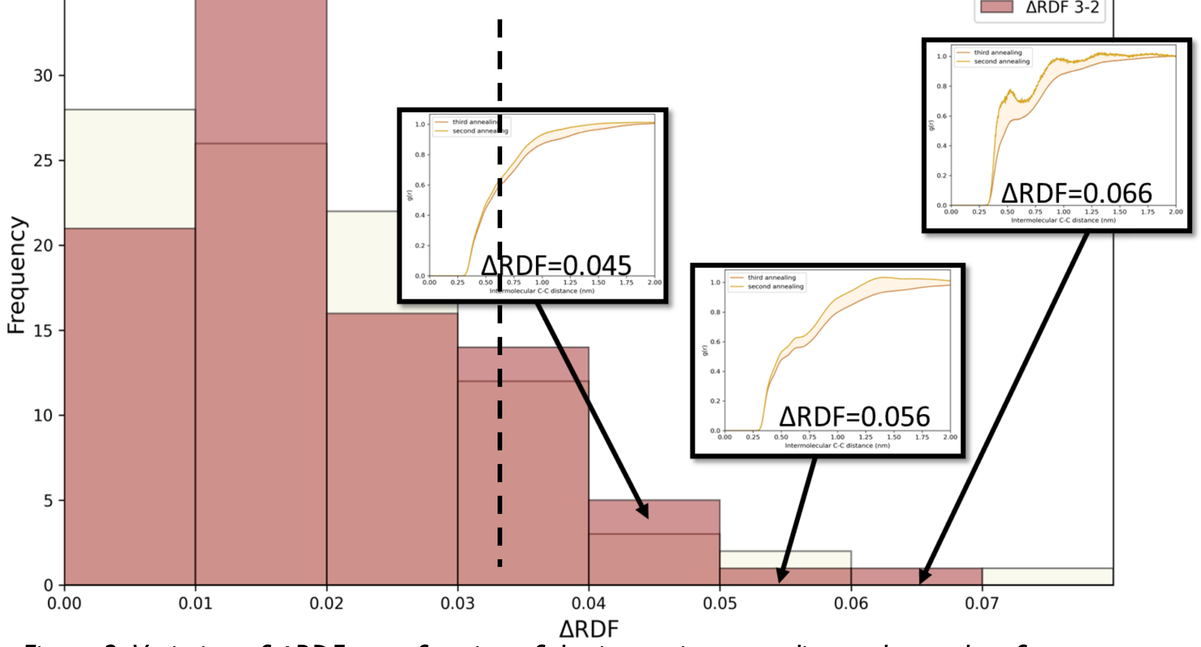
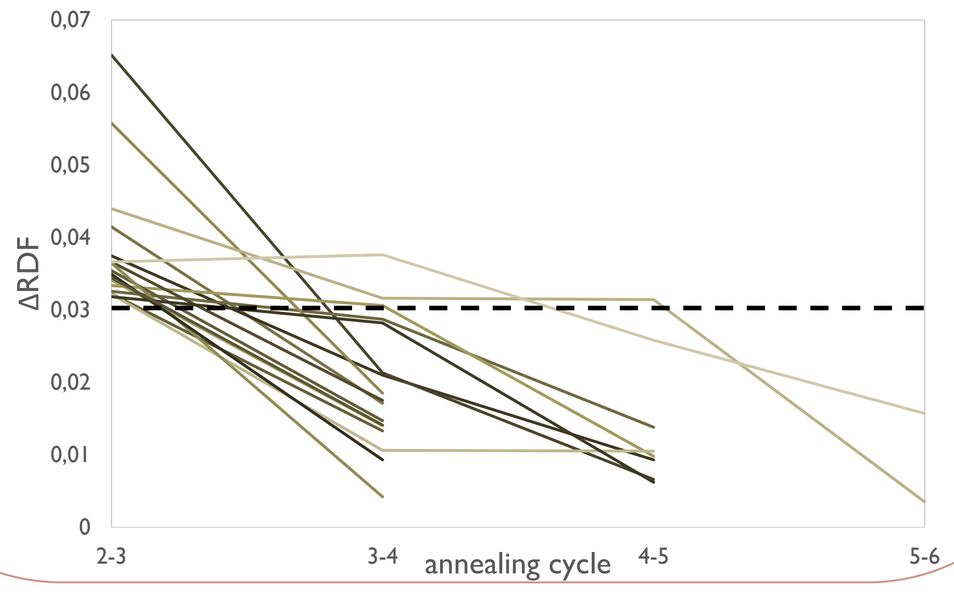


Figure 2. Variation of ΔRDF as a function of the increasing annealing cycle number, for system above the set threshold



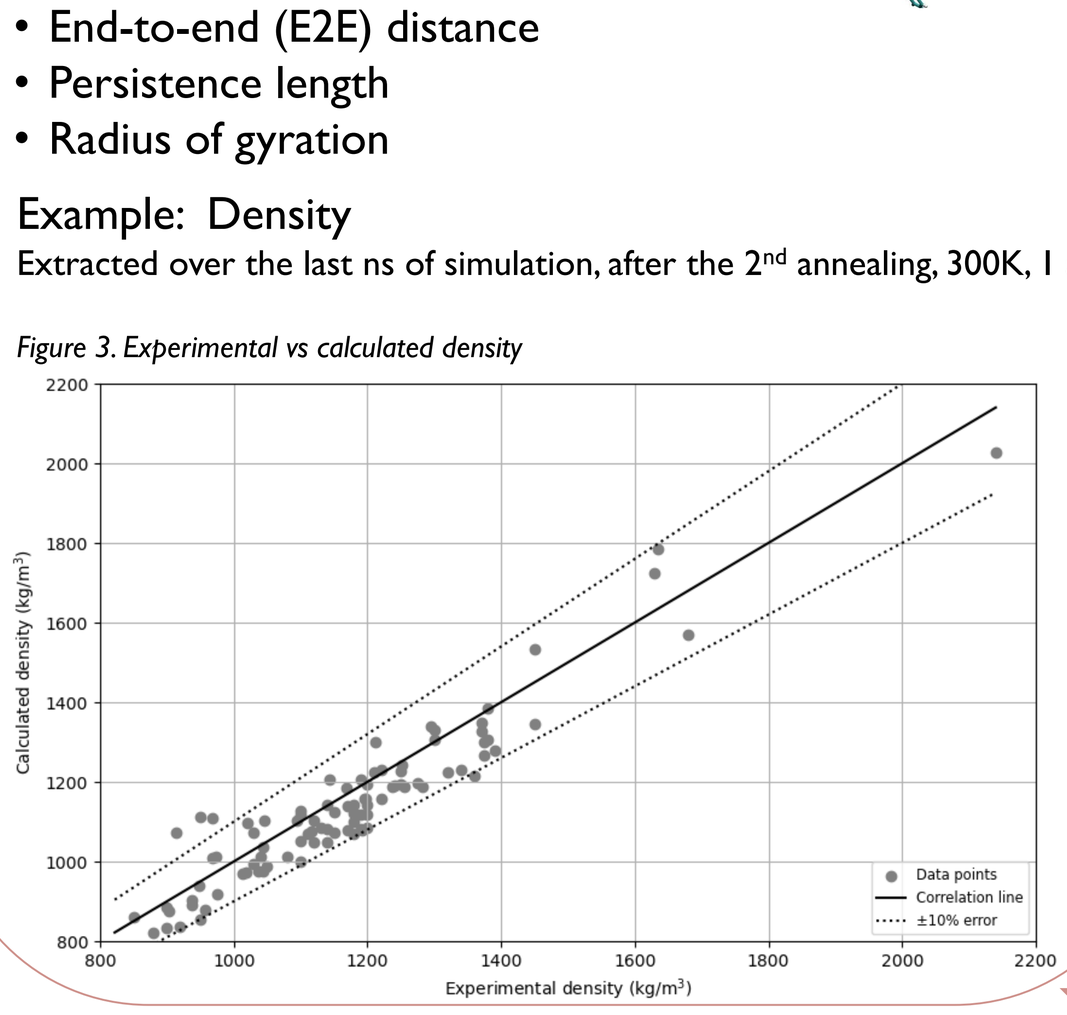
3- Analysis

- Density
- Contour length
- End-to-end (E2E) distance
- Persistence length
- Radius of gyration

Example: Density

Extracted over the last ns of simulation, after the 2nd annealing, 300K, 1 atm

Figure 3. Experimental vs calculated density



Machine Learning property prediction

Random forest models with 100 trees were trained using 5-fold cross-validation. MACCS fingerprints and structural features (e.g. contour length, end-to-end distance, density) were used as input to predict persistence length and density.

Figure 4. Predicted density vs experimental density

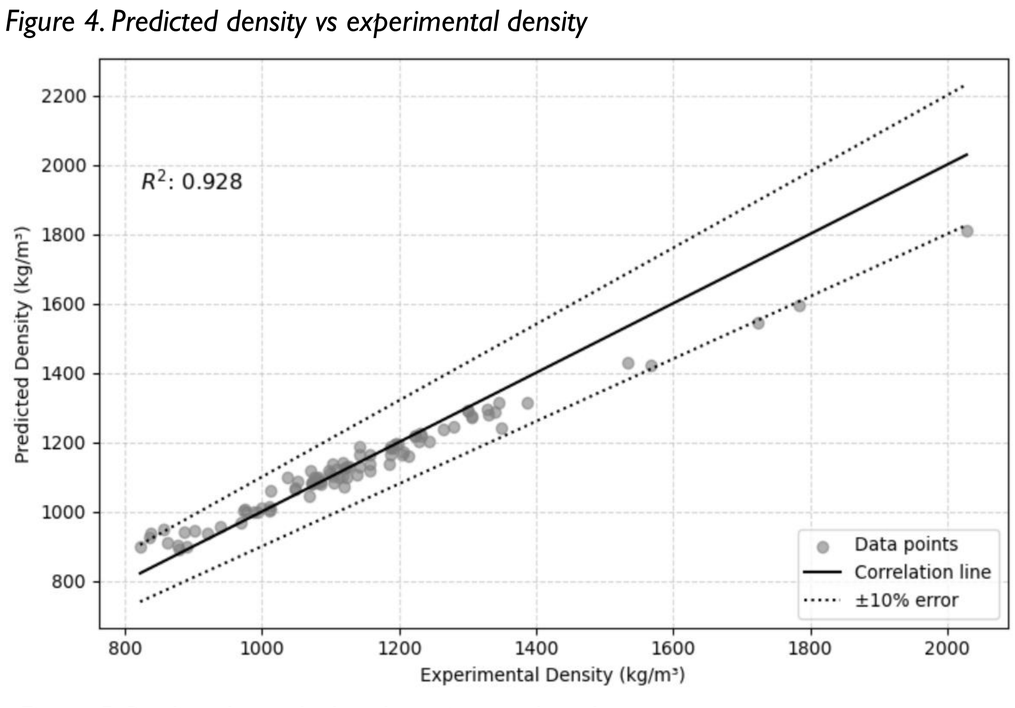
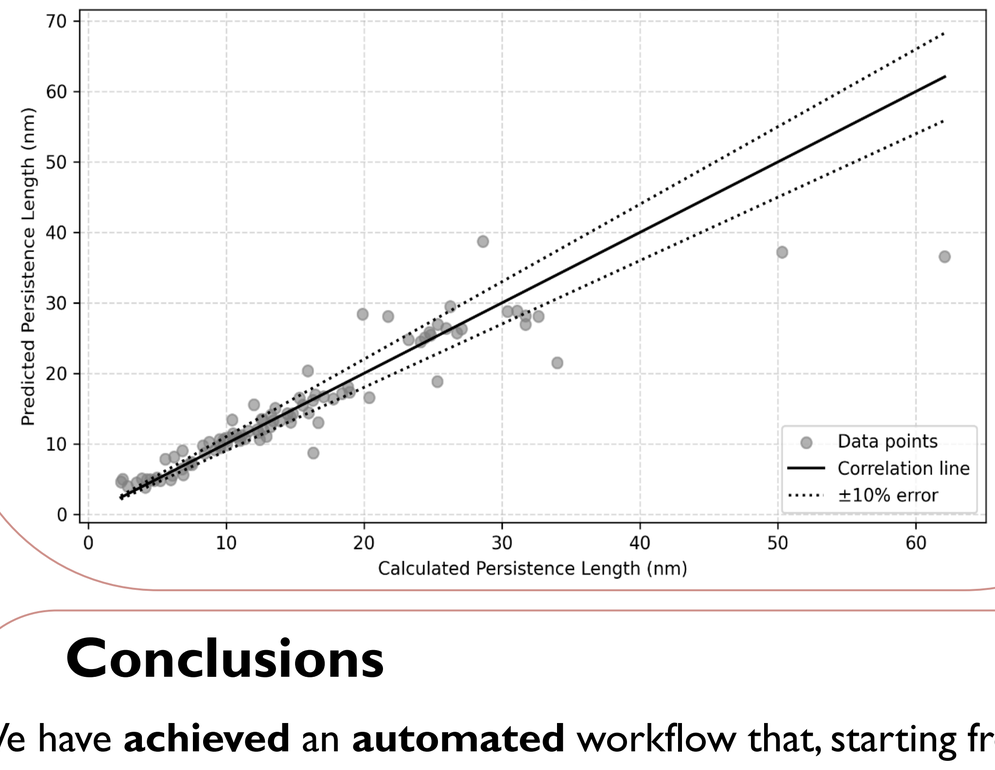


Figure 5. Predicted vs calculated persistence length



Conclusions

We have **achieved** an **automated** workflow that, starting from the SMILE of a monomer, returns the properties of a polymer

- **Fast**: topology and initial file generation 15 min human time
- **Cheap**: optimised simulation of 100 **equilibrated models** in about 150 hours per 4 CPU each
- **Homogeneous**: in terms of simulation timescale and condition
- **Versatile** protocol, tested on 100 polymers: only 1 crash
- **Reliable** in generating equilibrated structures, **validated** against available data

References

[1] *"Polymers: A Property Database" by B. Ellis and R. Smith, CRC Press, 2008*
[2] *J. Chem. Theory Comput. 2022, 18, 4, 2737–2748*
[3] *Nucleic Acids Res. 2017 Jul 3; 45(Web Server issue):W331–W336*
[4] *Breiman, L. (2001)., 45(1), 5–32*
[5] *Wu, S., Kondo, Y., Kakimoto, M. A., Yang, B., Yamada, H., Kuwajima, I., & Yoshida, R. (2019), 5(1), 66.*