Predicting the Glass Transition Temperature of Biopolymers via High-Throughput Molecular Dynamics Simulations and Machine Learning

Didac Martí,[†] Rémi Pétuya,[†] Emanuele Bosoni,[†] Anne-Claude Dublanchet,[‡] Stephan Mohr,[*,†] and Fabien Léonforte[*,‡]

[†]*Nextmol (Bytelab Solutions SL), Barcelona, Spain*

[‡]*L'Oréal Group, Research & Innovation, Aulnay-sous-Bois, France*

E-mail: stephan.mohr@nextmol.com; fabien.leonforte@loreal.com

Abstract

Nature has only provided us with a limited number of bio-based and biodegradable building blocks. Therefore, the fine tuning of the sustainable polymer properties is expected to be achieved through the control of the composition of bio-based copolymers for targeted applications such as cosmetics. Until now, the main approaches to alleviate the experimental efforts and accelerate the discovery of new polymers have relied on machine learning models trained on experimental data, which implies an enormous and difficult work in the compilation of data from heterogeneous sources. On the other hand, molecular dynamics simulations of polymers have shown that they can accurately capture the experimental trends for a series of properties. However, the combination of different ratios of monomers in copolymers can rapidly lead to a combinatorial explosion, preventing the investigation of all possibilities via molecular dynamics simulations.

In this work, we show that the combination of machine learning approaches and high-throughput molecular dynamics simulations permits to quickly and efficiently sample and characterize the relevant chemical design space for specific applications. Reliable simulation protocols have been implemented to evaluate the glass transition temperature of a series of 58 homopolymers, which exhibit a good agreement with experiments, and 488 copolymers. Overall, 2,184 simulations (4 replicas per polymer) were performed, for a total simulation time of 143.052 μs. These results, constituting a dataset of 546 polymers, have been used to train a machine learning model for the prediction of the MD-calculated glass transition temperature with a mean absolute error of 19.34 K and a R^2 score of 0.83. Overall, within its applicability domain, this machine learning model provides an impressive acceleration over molecular dynamics simulations: the glass transition temperature of thousands of polymers can be obtained within seconds, whereas it would have taken node-years to simulate them. This type of approach can be tuned to address different design spaces or different polymer properties and thus have the potential to accelerate the discovery of new polymers.

1 Introduction

Polymers are ubiquitous in our society thanks to their relative ease of synthesis from petroleum-derived intermediates.[1] Synthetic polymers can be formulated into diverse materials, have shown extreme durability, and can be manufactured at a low cost.[2] However, due to current environmental concerns, the replacement of fossil-based polymers with biopolymers has emerged as a promising mitigation strategy for reducing greenhouse gases emissions[3] and contributing to a more circular economy.[4] Plants and seafood waste are examples of renewable feedstocks containing natural biopolymers such as cellulose/hemicellulose, starch, lignin, and chitin, and have exhibited a broad range of structures, properties and functions in nature.[5] The current *in silico* study addresses the polymer structures and physicochemical properties from the monomer scale (in machine learning approaches) to the nanometer and

nanosecond scale (in molecular dynamics simulations). Thus, as this work cannot assess sustainability over the entire product life cycle, it is focused on, but is not limited to, bio-based polymers, hereafter referred to as "biopolymers". Indeed, despite their high interest and relevance, the questions of biodegradability and sustainability of biopolymers are beyond the scope of the current work. On the one hand, state-of-the-art procedures to assess biodegrability of chemicals, polymers and plastics are experimental.[6–8] On the other hand, the complex assessment of sustainability benefits and trade-offs requires scrutiny of the entire life cycle of polymers (feedstock harvesting, processing steps, and end-of-life scenarios). For more detail, the reader is referred to recent and thorough reviews.[4,5,9] Besides, while the terms polymers and plastics are often used interchangeably depending on the context, here we reserve the usage of plastics to address commercial products made from processed polymers while we focus our study on biopolymers for non-plastic applications.[4]

The cosmetic industry has always been at the forefront of innovation while staying closely connected to natural products.[10] It uses polymers for a variety of functions, e.g. as thickeners, conditioners, emulsifiers, film formers, etc.[11] Despite their extended application, the use of polymers is not unproblematic. Amongst others, there are concerns with respect to biodegradability, toxicity, and their sustainable production.[5] Apart from the underlying scientific and medical reasons, there is also an increasing regulatory pressure, which makes the development of new products and macromolecules that obey the standards of green and sustainable chemistry with sustainable ingredients a major goal of many industries.[4,12] Moreover, socioeconomic studies have shown that consumers perceptions of biodegradable polymers are positive and that they are willing to pay a higher price for green or sustainable products.[1,13] Therefore, there is nowadays an important drive to replace some of the typical polymers used in cosmetic products by alternatives that do not suffer the aforementioned shortcomings. However, in order to compete with established fossil fuel-based polymers and motivate their replacement, biopolymers must exhibit performance advantages harnessing the chemical functionalities of the starting bio-feedstocks,[14] and because of the complexity of

most of the formulas, solutions, emulsions, and other mixtures of components that industries use, the development of new, safe, and biodegradable products is a particularly challenging task.[15] Lots of efforts nowadays focus on making such formulas more and more natural by substituting historical compounds by naturally derived molecules, e.g., polysaccharides, with no trade-off on performances. However, these kinds of formulas are often multiphasic systems whose properties are highly sensitive to the substitution/addition of new compounds and most of the inherent mechanisms that are involved in systems under such conditions are still poorly known. In addition, the challenge is not only to correctly address physical or chemical rules or to provide knowledge to the science of formulation, but also to integrate a fair representation of biological substrates such as hair and skin. Such strategy has been applied recently for gathering knowledge on the modes of actions of performance engines, i.e., the set of common rules and components that drives an ensemble of formulations, for hair care applications.[16,17] However, such a fully digital evaluation workflow is well suited only when one knows which kind of molecules should be studied and evaluated, and for which purpose, i.e., the key performance levers have been clearly identified. For all these reasons, a strategy for addressing reformulation, substitution, and design of new ingredients consists in indirectly evaluating their cosmetics performance by targeting key features and physico-chemical levers that have been identified to impact the performance from data and experiments analysis. Therefore, in that context, the problematic takes the form of gathering relevant insights for subsequent formulation design by characterizing the polymers via a series of physico-chemical descriptors that can serve as proxies to key performance indicators.

To reach these ambitious objectives, the many major challenges to tackle will require significant and multidisciplinary R&D efforts.[5,18] Computational sciences, through simulation-based and data-driven approaches, will play a primary role to narrow down the search within the nearly infinite combinatorial polymer design space made of homopolymers, copolymers, and polymer blends of all sorts.[14,18] The development of tools capable of predicting the properties of polymers, both via molecular dynamics (MD) and machine learning (ML), without

resorting to the complete and heavy experimental procedure of synthesis and characterization, has the potential to dramatically accelerate the identification of promising candidates for targeted applications. Such approaches already appear as a cornerstone of the transition towards a greener and more circular economy, as they allow to study more chemicals in less time, provide deep atomistic insights, reduce laboratory-generated waste, and decrease the risk of R&D. Thanks to the important on-going effort undertaken in many fields of material sciences to collect, curate and provide access to both empirical and computational databases of materials and to the development of publicly accessible ML resources and tools, the design of new materials has already started to integrate data-driven approaches.[19,20] This is particularly attractive in polymer science, because, as soon as one starts combining monomers in copolymers, with different ratios and synthesis conditions, one is facing a combinatorial wall of possibilities. The Polymer Genome platform[21,22] is a bright example of a commercial endeavor that harnesses data spanning from the general polymer literature to density functional theory (DFT) calculations, in order to predict a series of polymer properties from the SMILES code[23] of the polymer. Some examples of these polymer properties are the glass transition temperature, polymer compatibility with 24 solvents, or polymer gas permeability. In order to extend their work to copolymers,[24,25] the same group collected more than 7,500 data points (which corresponds to about 40% of their final dataset) from the PoLyInfo database.[26] While others also report having collected large amount of data for more than 12,000 homopolymers from PoLyInfo,[27,28] MatNavi, the database service hosting PoLyInfo, now explicitly forbids the "acquisition of large amounts of data, whether by manual or mechanical means".[29] Therefore, as pointed out by Gormley et al.,[30] even though several polymer databases have been assembled, their data are not freely accessible in a downloadable and programmatic manner.[26,31–33] Consequently, a series of works have assembled disparate datasets for the prediction of a variety of properties, from the ionic conductivity of solid polymer electrolytes[34] to the Young modulus of polyurethane elastomers[35] or the performance as organic photovoltaic material.[36] More detail on the exciting progresses

of the polymer informatics community can be found in several recent review articles.[19,37,38] One of the difficulties in the development of polymer database resources is their stochastic and hierarchical structures with multiple length scales. This is a consequence of the statistical nature of the polymerization reactions that yield structures with distributions in molecular mass, composition, and topology. To overcome these barriers, a very recent initiative is impulsing the development of a Community Resource for Innovation in Polymer Technology (CRIPT), in particular with the objective of defining a scalable polymer material data structure.[39] Driven by FAIR (findable, accessible, interoperable, and reusable)[40,41] and open-source principles, CRIPT has the potential to accelerate the democratization and the adoption of polymer informatics. An alternative strategy, which permits to avoid the heavy workload of dealing with huge amounts of experimental polymer data, is possible, as evidenced by the release of RadonPy, an open-source Python library that automates polymer properties calculations using DFT and all-atom classical MD simulations.[42] In this work, Hayashi et al. successfully validated simulation protocols by systematically comparing the 15 calculated properties, which include thermal conductivity, bulk modulus and refractive index, amongst others, with experimental results for more than 1000 amorphous polymers from PoLyInfo. Furthermore, the recent software developments within the atomistic modeling community permit to harness the power of graphical processing units (GPU) hardware, which provides a tremendous acceleration of both classical and quantum-mechanical simulations. Therefore, RadonPy positions high-throughput (HT) simulations as an efficient and reliable data source. Polymer simulations are now capable of quantitatively predicting experiments, enabling the deployment of additional polymer informatics approaches and their rapid expansion to new regions of the polymer design space.

The glass transition temperature (T_g), defined as the temperature at which an amorphous material transitions from a glassy state into a rubbery state, is an important characteristic of polymers. For instance, when used as hair conditioning products, high T_g polymers are stiffer but also brittle, whereas low T_g polymer are more flexible at ambient temperature but

also provide less hold.[43] However, despite governing both manufacturing and applicability of the polymers, T_g is a notable absent from the list of RadonPy properties. In MD, the T_g can be calculated via annealing dynamics from high to low temperatures,[44,45] mimicking the approach that is applied experimentally. Accessing, via MD simulations, properties over a broad range of temperatures is challenging and computationally intensive, as it requires multiple equilibrations of the polymeric systems. Besides, there is a well known mismatch with respect to cooling rates, which are many orders of magnitude higher in MD simulations (nanosecond time scale) in comparison to experiments (100 seconds time scale).[46,47] For these reasons, many works have relied on machine learning approaches to predict the T_g of homopolymers[21,27,37,48,49] and copolymers[24,50] from experimental data. Nevertheless, Afzal *et al.* have evidenced the possibility to perform accurate HT calculations of the polymer T_g via MD, demonstrating good agreement with experiments.[51] Even though their study was limited to 315 homopolymers (note that they performed 10 replica simulations per homopolymer), which is not sufficient if one wants to perform a systematic screening of a broad chemical space that should also include complex copolymers, it shows great promises for *in silico* driven polymer design. Furthermore, Tao et al.[27] also obtained a consistent trend between MD-calculated T_g and experimental data for 100 homopolymers collected from PoLyInfo.

One great advantage of simulations is that, once the simulation protocols have been finely tuned and validated against experiments, they can be automated and their data and metadata are intrinsically machine readable.[38] Therefore, the combination of simulation- and data-driven approaches offers the flexibility and scalability to study hundreds to thousands of polymers in a fast and accurate way, as required for the exploration of new regions of the polymer design space. This is of high interest since approaches to predict T_g usually exhibit a trade-off between generality and predictive accuracy, as commented in detail by Pilania et al.[50] Indeed, as they illustrated for the polyhydroxyalkanoate polymers family, models designed to target a wide and diverse chemical space usually show higher predictive errors

and model uncertainties than models with a narrower applicability domain.

Throughout the work reported in this paper, MD simulations have been harnessed to acquire accurate and consistent data with the objective of generating tailor-made ML models targeting a specific region of the chemical space. We will show that this mixed approach, in comparison to generic models trained on experimental databases, permits to reach promising predictive accuracy with 10 to 20 times less data. More specifically, the goal of this work has been to train an ML model capable of predicting the T_g of biopolymers. High-throughput MD simulations of polymers have permitted us to obtain the T_g of 546 polymers, which constitutes the target property of our supervised ML model. The next section details the entire methodology implemented along this work. It addresses both the molecular modeling aspects and the data-driven strategy. Afterwards, the results obtained from the MD simulations are reported and compared to experimental measurements before presenting and discussing the performance of the ML model.

2 Methods

2.1 Molecular modelling

2.1.1 Challenges of setting up polymer simulations

Simulations of polymers are challenging due to the intrinsically large size of the involved macromolecules, which affects both the initial setup of the system and the actual MD simulation. The preparation of such systems requires the generation, in an automated but yet flexible way, of the configuration of the polymer, i.e., the coordinates of its constituent atoms, as well as the corresponding topology, i.e., the parameters describing the interactions of the constituent atoms according to the chosen force field (FF). Many tools have been recently developed to facilitate this task, and although an exhaustive review is out of the scope of

this paper, some specificity of a series of polymer builders permitting to generate all-atom systems for MD simulations have been summarized in the Supporting Information.

The first issue faced when building the topology is the assignation, in a consistent and automated way, of reliable force field parameters (i.e., bonds, angles, dihedrals, and non-bonded interactions) to the complete series of polymer structures to be constructed. A brief review of the FF generally used to determine the T_g of polymers via MD simulations is reported in the Supporting Information. While the investigation of a specific family of polymers often shows better agreement with experiments when relying on further refined FF parameters,[52] for example torsion potentials, such case-by-case refinements would be far too time consuming for an extensive study like the present work. Considering the diversity of polymer structures, aiming at reproducing the exact experimental values in a consistent HT manner is utopian. Instead, the main focus should be on correctly producing relative comparisons between different polymers. Therefore, the large but roughly constant offsets of CHARMM General Force Field (CGenFF)[53,54] and of the General Amber Force Field (GAFF)[55,56] for the estimation of T_g are not a problem,[52] as long as the relative ranking between polymers follows the experimental trend. For instance, even though their MD simulations overestimate T_g values by 79.1 K on average compared to experimental values, the results obtained by Afzal et al.[51] clearly reproduce the experimental trends, which is highly valuable in the perspective of simulation driven polymer design approaches. As our objective is to perform the HT study of a set of diverse biopolymers, we decided to use GAFF[55,56] for our simulations, both because it has been extensively used for a wide range of systems, demonstrating its good transferability and validity, and because its parameter assignation can be readily automated via Antechamber[56] or with a tool like ACPYPE.[57] Moreover, it has already been employed in several studies related to the T_g calculations. For instance, Alesadi et al.[58] obtained good agreement between MD simulations, experimental references, and an ML model for semiconducting conjugated polymers, and Andrews et al.[59] calculated several physico-chemical descriptors, among others T_g, for poly-lactic-co-glycolic acid (a synthetic

biodegradable copolymer), obtaining good agreement with available experimental references. As will be shown later, our simulations based on GAFF have indeed led to results that are in good agreement with similar simulation studies and experimental reference values. However, we have noticed that for a specific family of biopolymers, namely the polysaccharides, GAFF did not perform well and rapidly led to instabilities in the MD simulations when the temperature was increased. Therefore, in these cases, we have parameterized the corresponding monomers with the force field GLYCAM06.[60] GLYCAM06 was originally designed with the objective of introducing a minimal set of parameters required to add carbohydrate simulation functionality to the AMBER force field,[61,62] while maintaining consistency with that FF.[63] Since GAFF has also been developed consistently with AMBER, it is compatible with GLYCAM06, as the others FF from the AMBER family,[64] and it should be possible to combine them. In this context, parameter orthogonality is ensured assigning unique atom types for GLYCAM06, consistently with the Antechamber procedure. In practice, the appropriate assignation of GLYCAM06 atom types to polysaccharide structures relies on the recognition of atomic fragments by a homemade implementation analyzing the chemical environment of each atom.

In a second step, in order to obtain the complete topology, most modern FF require to assign atomic partial charges to all atoms of the system. This assignation is typically based on quantum mechanical calculations and a subsequent fitting of the atomic charges to reproduce the electrostatic potential,[65] or semi-empirical methods that were parameterized to produce atomic charges that emulate the same potential.[66,67] However, the computational time of this approach can become problematic for large systems, as it scales at least cubically with the system size. Even though there exist reduced scaling methods that make the underlying *ab initio* calculations more benign for very large systems,[68] they typically introduce extra overhead. In many recent polymer builders, whose capabilities and particularities are summarized in the Supporting Information, the charge assignation is often not explicitly discussed.[69–75] On the other hand, *Polyply*[76] uses parameterized charges from

GROMOS 2016H66 FF,[77] *Polymatic*[78] works with the charges provided by the user, *Polymer Structure Predictor*[79] uses monomer charges calculated on isolated monomers, and *BIOVIA Materials Studio*[80] permits to use FF-defined charges, charges obtained from Qeq,[81] or from Gasteiger[82] methodology. To account for the local environment of the monomer to some extent while keeping the computational cost manageable, oligomers of the homopolymeric systems were considered.Within this simplified setup, the atomic charges are calculated based on the AM1-BCC approach[66,67] and stored for each atom of the three different parts (head, repeat unit, tail).

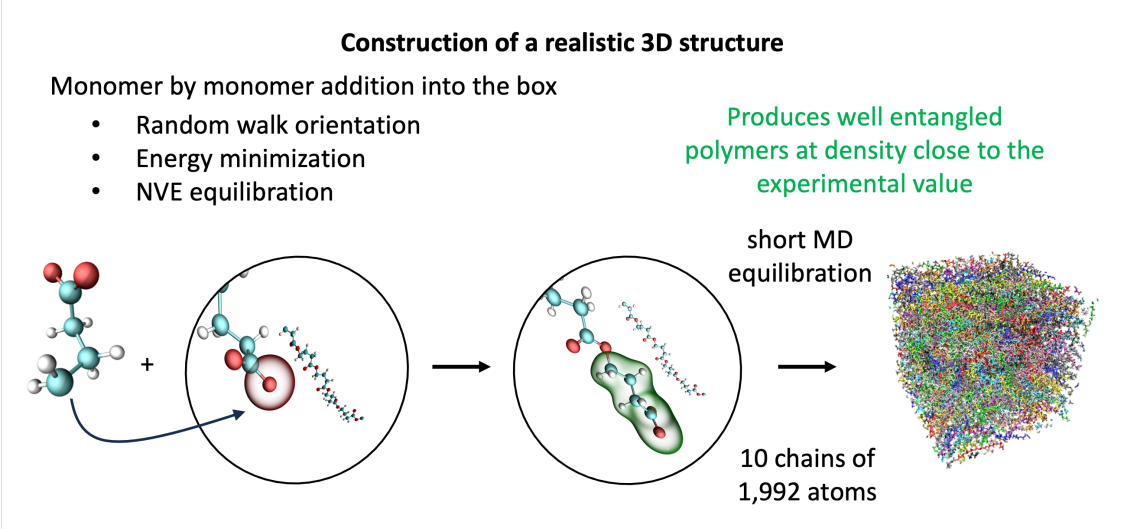

Figure 1: Summary of the building procedure implemented to obtain a well-entangled polymer melt structure before performing the T_g molecular dynamics simulations.

Once the relevant force field parameters are properly assigned and the partial atomic charges are obtained, the corresponding topology is constructed. Then, the next challenge is to set up a realistic three-dimensional structure of the polymer. Due to the large size of the polymers and their inherent long relaxation times, the system can easily end up trapped because of an unfortunate bad initial setup, thus failing to reach an equilibrium state within the accessible simulation time.[44] Escaping from such trapped configurations can be accelerated by Monte Carlo approaches, but their parametrization can be tedious and error prone.[83] Therefore, as displayed in Figure 1, it is important to carefully prepare an initial structure

close to its equilibrium state, which makes the subsequent simulations faster and more reliable. The different strategies adopted by recently published polymer builders are reviewed in the Supporting Information. In our work, the polymer is grown monomer by monomer via a random walk. Specifically, the monomers are added in the desired order in the simulation box, and a short energy minimization and MD equilibration (in the NVE ensemble) with LAMMPS[84] is performed after each addition. In our approach, these MD steps are performed with the relevant modern force fields GAFF or GLYCAM06, conversely to *PSP* and *PySoftK* that rely on former generic ones such as Universal Force Field (UFF)[85] and Merck Molecular Force Field (MMFF).[86] Once the coordinates of the entire chain of N monomers (1 head + $(N-2)$ repeat units + 1 tail) are obtained, the final polymer topology is generated using ACPYPE.[57] In this way, we can produce a system that is already well pre-equilibrated and ready to be used in the main MD simulations, only requiring a short equilibration and without having to resort to an elaborated yet cumbersome compression/relaxation procedure like the so-called 21-step procedure.[78,87] In order to illustrate this point, we compare in Figure S1 a snapshot of a system containing 10 chains of poly(4-hydroxybutyrate) (P4HB), each made of 77 monomers (1,992 atoms) created with this approach (bottom) to a system built with an approach in which each polymer is created independently and then added to the simulation box (top). As can be seen, the latter approach leads to a very badly mixed system of collapsed polymers, whereas our approach leads to a homogeneous mixture of entangled polymer chains.

2.1.2 Simulation setup

The simulations boxes are filled with 10 polymer chains of roughly 2,000 atoms each, according to the approach described above. The box size is chosen such that the final system has a density of $0.5\,\mathrm{g/cm^3}$, which leads to box sizes of the order of $7\,\mathrm{nm} \times 7\,\mathrm{nm} \times 7\,\mathrm{nm}$. Even though the polymer chains are grown at a density that is typically of the order of 40%-60% below the experimental one, their adequate entanglement allows to reach well-

equilibrated systems with correct densities after relatively short simulation times in the NPT ensemble without resorting to complex multi-step equilibration protocols[78,87–89] often used in the literature[27,49,90] and accurate results can be obtained with relatively simple workflows, as detailed in the Supporting Information and in Figure 2 (top panel). The chain by chain addition approach reported in Figure S1, on the other hand, yields system that are typically stuck at densities about 15% below the experimental ones, and would most probably require to follow more elaborated equilibration protocols to escape from such configurations.

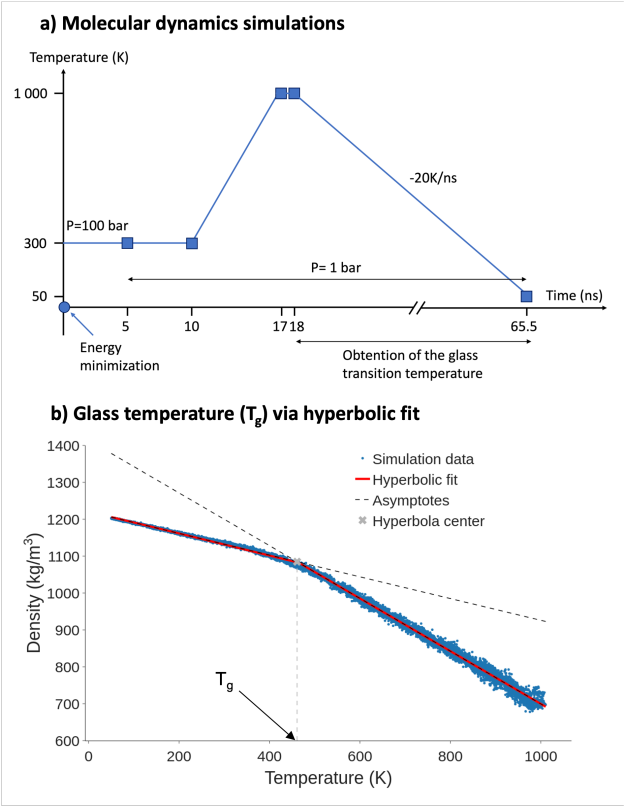

Figure 2: Simulation protocol implemented for obtaining the glass transition temperature of polymers (top) and illustration of the hyperbolic fitting procedure (bottom).

2.1.3 Glass transition temperature

In MD, the T_g can be calculated via annealing dynamics.[44,45] More specifically, its determination is based on the observation that, asymptotically, the density varies linearly with respect to temperature in both the glassy and the rubbery regimes, however with different

slopes; the T_g is then given as the temperature where the density slope transitions from the rubbery value to the glassy value. In practice, two linear functions can be fitted to the low and high temperature domains, respectively, and the T_g is then given as the temperature where they intersect.[91] Even though this approach is relatively straightforward to implement, it contains some inherent uncertainty as the transition between the slopes is not very sharp, and the specific value depends considerably on the exact definition of the low and high temperature regions and the corresponding fits. An alternative solution that eliminates this bias is to determine the T_g from a hyperbolic fit, as suggested by Patrone et al.[45] and illustrated in Figure 2 (bottom panel). Specifically, all the calculated data points are fitted to the following equation for the density as a function of temperature:

$$\rho(T) = \rho_0 - a(T - T_0) - b\left(\frac{1}{2}(T - T_0) + \sqrt{\frac{(T - T_0)^2}{4} + \exp(c)}\right), \quad (1)$$

where T_0, ρ_0, a, b and c are fitting parameters.[92] This approach is much better suited for systematic HT calculations, as there is no need to manually assign the data to the low and high temperature regimes, and we consequently employed it in this study for the determination of the T_g values.

In practice, we applied the simulation protocol summarized in Figure 2 (top) and detailed in the Supporting Information. It is worth mentioning that due to the inherent stochastic nature of the way our polymer systems are created, we used several replicas in order to increase the statistical significance of the obtained results. In the literature there are examples where 10 replicas[51] and 5 replicas[52] were used. In our case we have seen that using 4 replicas is enough to yield meaningful results with generally small standard deviations, as further shown in section 3, while limiting the overall computational cost of the entire study.

2.2 Data-driven approach

2.2.1 Dataset collection and combinatorial generation of copolymers

To assemble a dataset of monomers with a focus on biopolymers we have reviewed the literature on biopolymers and the freely accessible polymer databases.[4,5,9,32,93] As discussed in the introduction, this manual data collection is a cumbersome process. Besides, as the list of gathered known biopolymers, which includes polyhydroxylkanoates and polysaccharides, could be considered limited, we have expanded our dataset with a few common monomers such as polyvinyls, polyacrylates, and polyesters to reach 58 monomers. This number looks very small in comparison to the more than 12,000 monomers of PoLyInfo, but as will be discussed in section 2.2.3, they cover a smaller part of the polymer chemical space than PoLyInfo. To permit the exploration and to densify the sampling of this region of interest for our application, namely the ML prediction of the T_g of biopolymers, a set of 14,877 combinatorial copolymers were generated from the binary combinations of the 58 collected monomers with monomer ratios in the range [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. The list and SMILES code of the monomers used in this work is provided in the Supporting Information.

2.2.2 Encoding polymer structures for machine learning

For training an ML model to predict a polymer property, it is first necessary to encode the polymer structures into a set of features from which the algorithm will learn to make predictions. In our dataset, homopolymers are represented by the polymer SMILES (Simplified Molecular Input Line Entry System) code of their repeat unit,[23] in which the symbol "*" indicates the polymerization point. For binary copolymers, the SMILES codes of the two repeat units involved and their respective weight ratios are used. One might wonder if relying merely on repeat units SMILES would not limit the encoding of the complex polymer hierarchical structures, since alternative strategies have been tested or are currently under

development. Nevertheless, according to the state-of-the-art review of the alternative strategies reported in the Supporting Information and the very complete work from Tao et al.,[27] the repeat unit structure sufficiently captures the bonding information and key substructures for the ML prediction of T_g of polymers. Therefore, in this work, we have encoded the polymer structures using molecular descriptors calculated from the polymer repeat units with the open access Python package Mordred.[94] In particular, we have used the 1,613 2-dimensional (2D) molecular descriptors that Mordred calculates. For the hypothetical copolymers generated from the binary combination of monomers, the feature vectors have been generated as linear combinations of the descriptors of the involved repeat units weighted by their corresponding ratios, following the same strategy as recent studies by Pilania at al. and Kuenneth et al.[24,25]

2.2.3 Selection of polymers to simulate

As previously discussed, the existence of a trade-off between generality and predictive accuracy for ML models developed to predict the T_g of polymers has been observed in several occasions.[50] Therefore, to select the polymers worth simulating as efficiently as possible, it seemed of interest to first position the homopolymers and hypothetical copolymers in the broader context of polymer chemical diversity. Despite the limitations to access the broad information of polymer databases presented in the Introduction, the work of Ma et al.[28] released a useful benchmark database for polymer informatics. Training a generative model on 12,000 homopolymers from PoLyInfo, they built a database of about 1 million hypothetical polymers named PI1M. Moreover, they showed that PI1M covers a similar chemical space as PoLyInfo, while populating more densely regions where PoLyInfo data are sparse. Therefore, we have used PI1M as a proxy for the chemical diversity of synthetic polymers and implemented a selection strategy, detailed in the Supporting Information, which relies on dimensionality reduction and clustering approaches to sample easily and broadly the chemical space of interest for our application. Overall, we simulated all 58 homopolymers

built from the monomers collected and we identified 1,000 copolymer combinations via the approach detailed in the Supporting Information as upper dataset limit. Then, from this pool of 1,000 copolymer combinations, we randomly selected structures to simulate and to fill our dataset until the subsequent ML models reached a satisfying performance level. All polymers simulated are displayed in Figure 3 after a UMAP (Uniform Manifold Approximation and Projection)[95] dimension reduction; in green for the 58 homopolymers and in black for the 488 copolymers, while Figure S4 also shows all combinatorial copolymers. The only precaution that we took was to ensure that we did not over-sample combinations representative of little populated clusters positioned in the sparser and outer regions. Indeed, despite the non-linearity of the UMAP transformation, they would likely carry a higher probability of sampling sub-areas ultimately assessed as outside of the applicability domain of the ML model (to be defined at a later stage).

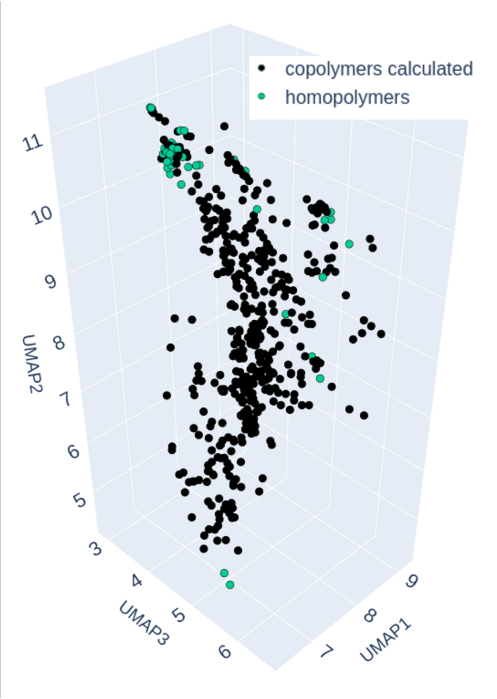

Figure 3: Representation of the homopolymers (green) and copolymers (black) constituting our dataset within the UMAP transformation trained on PI1M. Figure S4 offers an alternative orientation also displaying all generated copolymers.

2.2.4 Supervised learning

The final objective of this work is to train an ML model that is accurate enough to accelerate the exploration of the polymer design space for the targeted application via the prediction of T_g. The MD-calculated T_g of homopolymers and copolymers selected in the previous section is the target of the models, while polymer structures encoded as described in section 2.2.2 constitute the features. All models were implemented using scikit-learn library version 1.2.0.[96] The different training steps involed, their outcomes and the technical details are reported in the Supporting Information.

3 Results and discussion

3.1 Validation of the simulation protocols

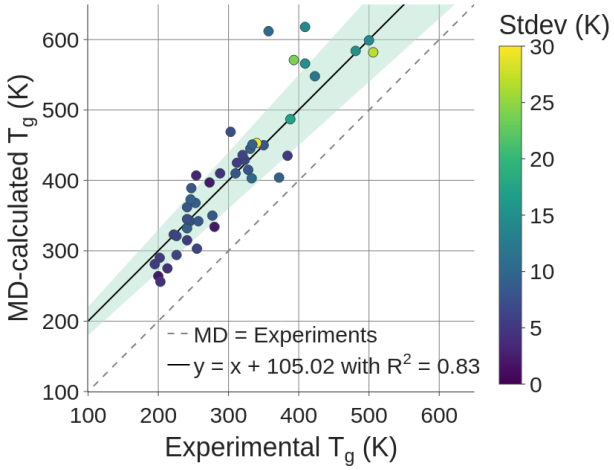

Figure 4: MD-calculated T_g against reported experimental values for homopoymers. The points are colored according to the standard deviation (Stdev) of the MD simulations, since for each system 4 replicas are performed.

Before undertaking the computationally demanding HT simulations for acquiring the T_g of copolymers, we have assessed the quality of the simulation protocols implemented via comparison of the MD-calculated T_g with available experimental results for the homopoly-

mers. In Figure 4, we show the T_g of 47 homopolymers obtained via the MD simulations against their experimental reference values. All these data, with the sources of the experimental references, which are either collected from the literature or from databases such as PoLyInfo[26] or the Polymer Database,[32] are provided in the Supporting Information. For the remaining 11 homopolymers we could not find any reliable experimental T_g reference, therefore they have been discarded from this comparison. In the case of multiple references being available in databases, we followed the prescription of Jha et al.[97] and took the median value of the entry. The MD-calculated T_g values are typically higher compared to their experimental counterparts, which is expected, as discussed in the introduction, due to the difference between the cooling rates used in the simulations compared to the ones employed in experiments. However, and importantly, the R^2 value of 0.83 evidences the preservation of the main experimental trends, i.e., relative comparisons between different polymers yield correct results, and this offset does not affect the usefulness of the obtained MD data. Indeed, the shaded area in Figure 4 has been drawn to show that most of the values are within 10% deviation from a simple 105.02 K offset from experimental references. Furthermore, even though we do not study the same polymers as Afzal et al.,[51] we note that accounting for the simple constant offset of our values while comparing MD-calculated T_g with experimental data provides an R^2 score close to what they obtain with linear fit. In other words, with increasing T_g, their MD results deviate from a constant offset while ours do not. Although some of our points Figure 4 seem to visually indicate a similar deviation, a linear fit of our data does not capture it, as will be discussed in section 3.4, and such effect seems very dependent of the dataset studied. Besides, the standard deviations on T_g values from our MD simulations seems smaller (see Figure S5 for histogram). Furthermore, it is worth mentioning that the homopolymers with the largest deviations from the experimental trend, namely pectine (357 K in experiment *vs* 612 K in MD), hyaluronic acid (409 K *vs* 618 K), and cyamopsis-tetragonobola (393 K *vs* 571 K), are polysaccharide polymers for which the experimental assessment of T_g is usually challenging and for which only limited references

could be found. Overall, the good agreement between simulations and experiments validates our approach to use MD simulations to sample efficiently and accurately this region of the polymer chemical space for the training of ML models.

3.2 High-throughput polymer simulations

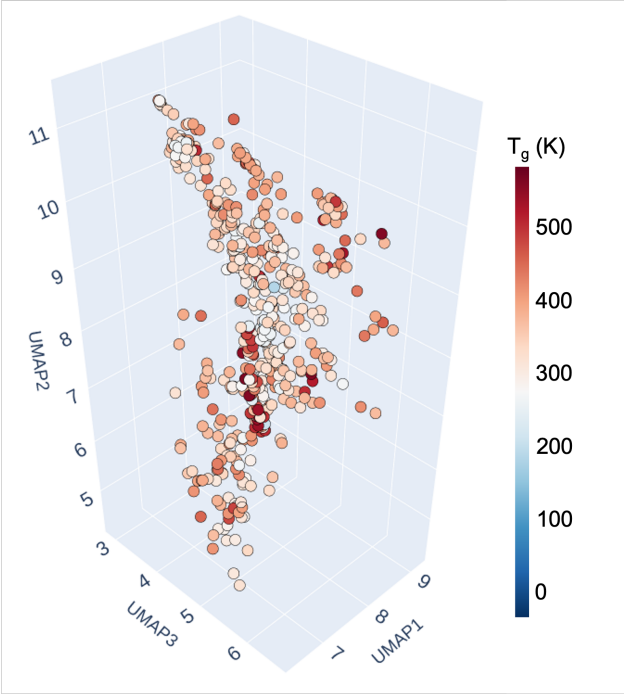

Figure 5: Representation of the chemical space of all homopolymers and copolymers sampled in this work, colored according to their T_g. The colored scale is centered in 273.15 K

Once our simulation protocol had been validated, the T_g of the 488 copolymers selected via the approach detailed section 2.2.3 was obtained by HT simulations. A total of 2,184 simulations ([58 homopolymers + 488 copolymers] × 4 replicas) were performed, which corresponds to a total simulation time of 143.052 μs for this entire study. In Figure 5, the chemical space sampled via MD simulations is displayed with each polymer colored according to its mean T_g value. Before exploiting the MD-calculated T_g as targets for the training of ML models, it is interesting to have a look at the generated data. As illustrated by the histograms in Figure S5, the standard deviation of the T_g values is generally very small. For homopolymers, the average standard deviation is 9.60 K, whereas for copolymers is is 6.70 K,

with only a few values above 20 K in both cases. The composition dependence of copolymers T_g is often rationalized by the Fox equation,[98] which states that the T_g of a copolymer can be derived from the homopolymer T_g values of its constituting monomers as follows:

$$\frac{1}{T_{g,mix}} \approx \sum_i \frac{\omega_i}{T_{g,i}}, \tag{2}$$

where $T_{g,mix}$ and $T_{g,i}$ are the glass transition temperatures of the copolymer and of the homopolymers of the constituting components, respectively, and ω_i is the mass fraction of component i. Despite the establishment of this relation more than 70 years ago, the investigation of the composition dependence of T_g is still very relevant, in particular for understanding the deviations from linearity[99] or the effects of monomers sequence.[100] Figure 6 compares the combination of the MD-calculated T_g values of homopolymers according to the empirical FOX relation with the actual MD simulation results for the corresponding copolymers. As expected, the majority of the Fox equation values (for 339 copolymers) lie within a small deviation of ±19.51 K from MD results, with an average deviation of 8.20 K for this subset of copolymers. However, the remaining 149 copolymers show a much larger average deviation of 45.23 K. Histograms of the deviation distributions for both cases are represented in Figure S6. If one were tempted, after assuming the effort required to obtain the T_g of a series of homopolymers via MD simulations, to use the Fox equation to predict the T_g of copolymers, a mean absolute deviation of 19.51 K could seem very satisfying. Nevertheless, resorting to such strategy would consist in erroneously positioning an empirical relation as a gold standard and treat any deviation as prediction error whereas these deviations from linearity, which are observed experimentally and currently rationalized as the consequence of intra-chain stiffness and polar interactions effects,[99] can be captured via MD simulations. Therefore, the implementation of an approach based on a combination of explicit simulations of copolymers and ML is better suited for the exploration of the relevant polymer design space.

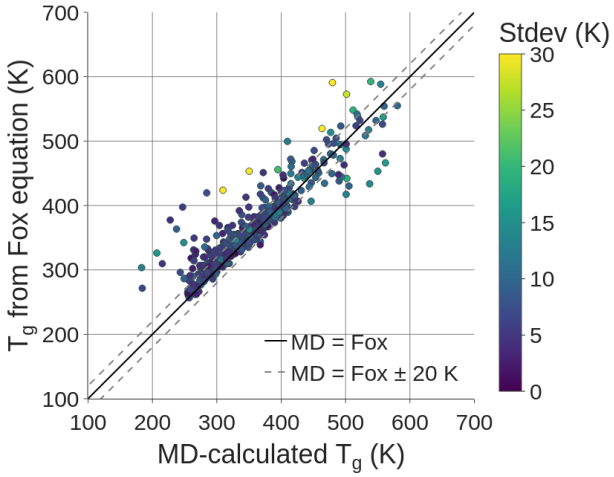

Figure 6: T_g values predicted using the empirical Fox equation[98] versus MD-calculated values for copolymers. The points are colored according to the standard deviation (Stdev) of the MD simulations, since for each system 4 replicas are performed.

3.3 Training machine learning models

After the usual separation of the data collected from MD simulations into a training dataset with 491 samples and a testing dataset with 55 samples via a 90/10 train/test split and the preprocessing of the training dataset, a series of regression algorithms (see Supporting Information for all technical details), including logistic regression, k-nearest neighbors, support vector machines, and ensemble learning algorithms, were compared for the prediction of the MD-calculated T_g of polymers. The mean absolute error (MAE) is used as scoring function throughout this work. The results obtained across the range of number of tested features (between 5 and 40) are displayed in Figure S7, and highlight the good performances of out-of-the-box random forest (RF), gradient boosting (GB), and k-nearest neighbors (KNN) regression algorithms, which were thus selected to be further refined and tested. Then, the best features to train the models have been identified in two steps: i) first, the 50 most relevant features were identified via recursive feature elimination (RFE), and ii) the effect of

the number of features on the model performance was investigated in detail within a leave one out cross validation (LOOCV), still relying on RFE for the features selection. It is worth mentioning that the RFE procedure requires to be implemented with an estimator, i.e., a regression model in this case, able to return a measure of features importances. Among the 3 algorithms selected, only RF satisfies this condition; thus, in order to perform its selection, the RFE procedure has been set up to rely on a standard RF regressor with default parameters except for the number of trees (n_estimators) which was limited to 50, as we intend to reduce trees complexity in subsequent steps. Figure S8 shows the evolution of both the MAE and the root mean square error of these evaluations as a function of an increasing number of features and evidences a first plateau in the performance improvement between 25 and 33 features in MAE, our main scoring function. Therefore, in the center of this interval, we selected the 29 features identified by the RFE procedure for all subsequent training steps with all learning algorithms. The list and definition of these features is reported in the Supporting Information. Afterwards, for each type of algorithm, the hyperparameters of the models have been optimized via grid searches detailed in the Supporting Information, in which the ranges explored by the parameters governing the complexity of the model, such as the number of trees (for RF and GB) or the number of neighbors (for KNN) used in the training were limited to prevent overfitting.

The learning curves represented in Figure 7 for the best KNN model and in Figure S9 for the best models obtained with each learning algorithm permit to evaluate the models average performances on training datasets and on hold out validation datasets, separated via cross validation (CV) procedure, for different sizes of training datasets. That task was performed using a 10-fold ShuffleSplit CV. On the one hand, the three models reach similar levels of performances when the entire training dataset is used, with MAE values around 20 K and the MAE curves obtained on the validation datasets keep decreasing (they do not plateau) at the maximum training set size used, indicating that the performances of the models could be improved by further addition of samples to the training dataset. This is

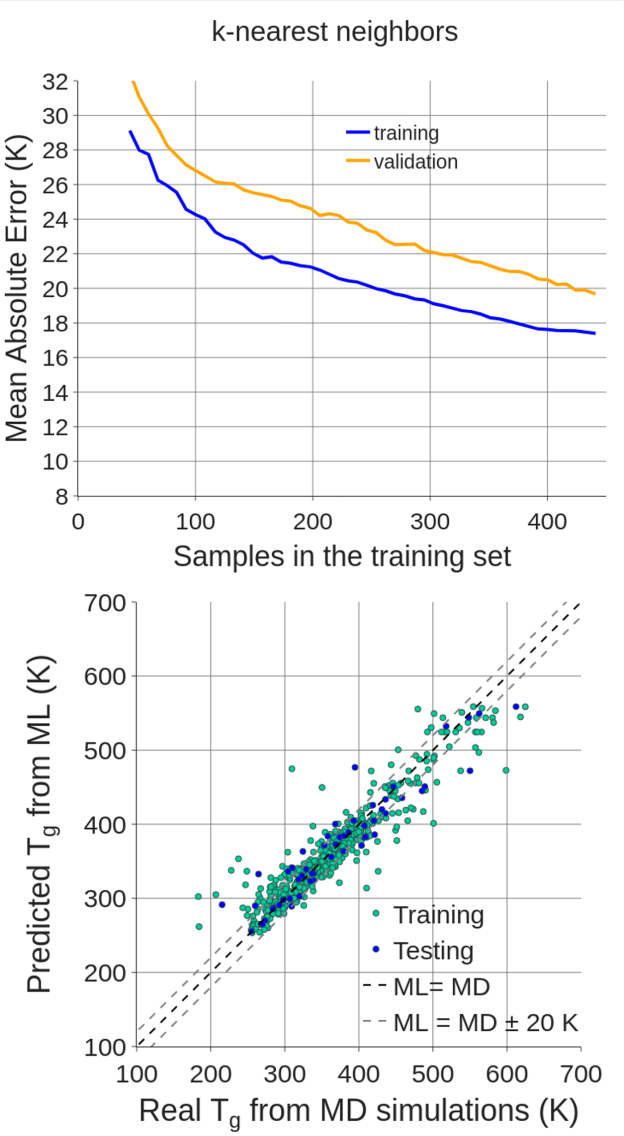

Figure 7: Top panel: learning curves on the training and validation datasets evaluated through ShuffleSplit CV as function of the size of the training datasets for the best KNN model obtained. Bottom panel: final performances of the best KNN model on training (green) and testing (blue) datasets.

a positive sign for model health and it further illustrates the great potential of the entire strategy implemented in this work. On the other hand, the RF and GB models exhibit a significantly higher variance than the KNN model, i.e., there is a gap between their learning curves on the training and on the validation datasets. The superior performances of RF and GB on the training datasets compared to the validation datasets indicates a slight overfitting, as also observed by Tao et al.[27] who reported similar MAE differences of the order of 10 K between training and testing dataset. Nevertheless, as the performances on the validation datasets are similar to the KNN model, we continued to characterize the performance of all three models and evaluated their generalization abilities both in 10-fold ShuffleSplit CV and in LOOCV, as reported in Table S2.

Table 1: Final performances on training and testing datasets of the best models obtained with each learning algorithm.

	Random Forest		Gradient Boosting		k-nearest neighbors	
	Testing	Training	Testing	Training	Testing	Training
Mean Absolute Error (K)	19.97	12.43	20.30	12.04	19.34	17.04
Root Mean Square Error (K)	29.81	19.54	28.30	16.61	27.98	26.56
R^2	0.88	0.93	0.89	0.95	0.89	0.87

The final performances of the best models trained with each learning algorithm are assessed by applying them to the testing set isolated before the start of the training procedure. Table 1 summarizes the results and shows that the KNN model is performing best with a MAE (the optimized scoring metric) of 19.34 K on the testing set, which corresponds to a very good coefficient of determination $R^2 = 0.89$. The good performances of the final model, illustrated in the bottom panel of Figure 7 that shows the T_g predicted by the KNN model as function of the reference MD-calculated T_g and permits to visually compare the performance of the model between training and testing datasets, are consistent with the evaluations of the cross-validations reported in Table S2. Furthermore, it is interesting to note that the good fit of the KNN model, which predicts the T_g of polymers based on the values of the neighboring structures in the chemical space, is consistent with the generally

relatively smooth transitions over domains of different T_g ranges visible in Figure 5. Our model performances are in the same order of magnitude as the work of Tao et al.,[27] which used 10 times more data to cover a larger design space. Their best models showed an $R^2 = 0.82$ (their main scoring metric) and a MAE of 34.45 K. With an entirely different approach, namely a meta learner learning from 5 previously trained cross-validation models relying on a total of 18,445 data points (20% for the meta-learner and 80% for the cross-validation models) for homopolymers and copolymers glass transition temperature, melting temperature, and degradation temperature, Kuenneth et al.[24] reached a root mean squared error (RMSE) of 21.03 K for $R^2 = 0.96$ on the capability of the meta learner to predict T_g for the 80% of data used in the cross-validation models. One might note that they do not report the isolation of a testing set to be excluded from any training step and only to be used for final performance evaluation. Furthermore, as they obtain an important performance improvement from their meta learner, we can point out that the RMSE of our model is of the same order of magnitude as the ones of their cross-validation models for homopolymers while their cross-validation models for copolymers perform better In these two other works, an impressive amount of experimental data is used.

Despite their slight overfitting on the training set, both the GB and RF models reported in Table 1 and illustrated in Figure S10 are performing well. Therefore, it seems worth having a look at the features importances of the final RF model shown in Figure S11 to put them in perspective with the glass transition process. Indeed, the shifts in polymer properties occurring when the polymers transition from the glassy state to the rubbery state are generally attributed to a change in molecular motion.[51] Consistently with this observation, the highest ranked feature for the RF model is the rotational ratio (ratio between number of rotatable bonds and number of bonds, discarding hydrogen atoms), which is a proxy for monomer flexibility and thus permits to capture differences in polymer motion due to their intrinsic monomer flexibility. Although they may not always be as intuitive as the rotational ratio, also other features ranked relatively high. On the one hand, AATSC0d

and ATSC4d encode the autocorrelation of sigma electrons at respectively very short and medium ranges; on the other hand, JGT10 and JGI2 describe the charge distribution at both long and short range. We interpret these range segmentations, also observed in lower ranked features, as ways for the RF model to encode changes in monomers properties at various scales. Short and medium range autocorrelation descriptors can capture property changes between neighboring atoms within one monomer in comparison to another one, for example to capture main chain/side chain interactions, which are also relevant for interpreting T_g variations,[99] whereas long range autocorrelation descriptors account for the differences over larger scale of the monomers. In the case of JGT10, this long range descriptor is a global mean from order 0 to 10, i.e., from topological distance (number of involved bonds of the shortest path between two atoms) 0 to 10, and thus provides an averaged encoding over the scale of the monomer (approximatively, since monomer size varies). In other cases, like with ATSC8d, the descriptor encodes the property at a large topological distance, and thus can capture differences between small and large monomers and between monomers with similar structures within the first neighbors of each atom, but with changes over larger distances. Such information is used by the RF model in its prediction process, but also by the other ML models, since they were trained with the same features. The last descriptors with feature importance superior to 0.05 is SssCH2, an electrotopological state indice summing the number of aliphatic -CH_2- carbons over the monomers. To gain further insights on the underlying machinery of the ML model predictions, one should turn to interpretable ML approaches, which we intend to explore in-depth in future works.[101]

3.4 Applicability of the final model

With the development of quantitative structure activity relationship (QSAR) models for regulatory usage on risk assessment of chemicals for their safe use, a series of principles have been established by the Organization for Economic Co-operation and Development (OECD) for the validation of QSAR models.[102] According to OECD guidelines, it is compulsory for

these models to have: i) a defined endpoint (in our case MD-calculated T_g of polymers), ii) an unambiguous algorithm (k-nearest neighbors with the specific parameters reported in the Supporting Information), iii) a defined domain of applicability, and iv) appropriate measures of goodness-of-fit, robustness and predictivity (as reported in the previous section). Although the context of our work is different, these principles represent standards to uphold. Therefore, the remaining task is to define the applicability domain (AD) of our final model, which defines the subspace of polymer structures for which the predictions of the MD-calculated T_g would be considered reliable. A large variety of approaches has been used for this task, from range-based and geometric methods to similarity-based methods and decision forests.[103,104] We have adopted a probability density distribution-based approach, namely the clustering-based local outlier factor (CBLOF),[105] which builds on the local outlier factor[106] at the cluster level. Therefore, using the implementation of the Python package PyOD,[107] we trained a CBLOF model to detect 15% of outliers within the dataset of 491 polymers used to train the ML model, from a Kmeans clustering model set to identify 20 clusters. In this procedure, only the 29 features kept as relevant for the final ML model were used as features for the outliers detection and the other parameters were kept at their default values. Using the web application hosted by Nextmol (see Supporting Information for access details), the reader can seamlessly apply the final KNN model trained in this work to obtain a prediction of the MD-calculated T_g for homopolymers and binary copolymers while obtaining an AD evaluation (is the polymer of interest within the AD of the ML model?) by simply providing the relevant SMILES codes and monomers ratios. Furthermore, to illustrate the acceleration provided by the approach implemented in this work for the exploration of the chemical space covered by the assembled dataset, we have applied the KNN model to predict the MD-calcualted T_g of the 9,029 combinatorial copolymers within its applicability domain. With the HPC architecture used in this work (the compute node contains 4 Nvidia A100 GPUs and 64 CPU cores), we are able to simulate 1936.0 ns per day. Therefore, as 4 replicas are needed, obtaining the T_g of these 9,029 combinatorial copolymers would

require 3.34 node − years, whereas the entire series of ML predictions is obtained within a few seconds. These predictions are displayed with the 496 MD simulations results in Figure S12 and Figure S13.

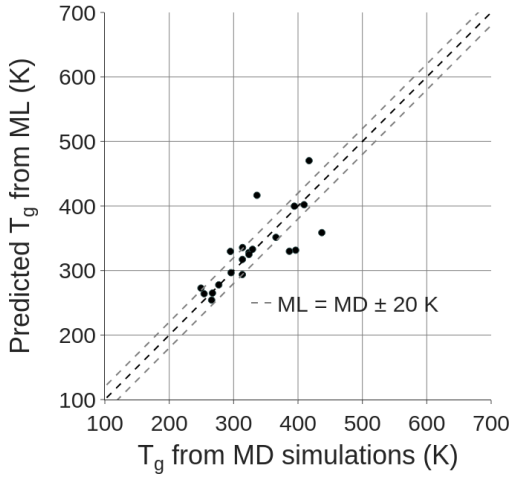

Figure 8: Performances of the final KNN model for the 14 additional simulated polymers within its applicability domain.

Moreover, as final evaluation of the approach implemented throughout this work, we have selected 21 polymers (listed in the Supporting Information) from the dataset of Afzal et al.[51] that are located within the AD of our ML model. Then, both the MD simulation protocols presented earlier and the final ML model have been applied to obtain the T_g of these homopolymers at both MD and ML levels. First, we compare in Figure S14 the results of our MD simulations with the results from Afzal et al.[51] and the experimental data that they collected from Bicerano.[108] Please note that in what follows, the experimental data are taken as reported, whereas comparisons with more recent works and revised experimental protocols may lead to different reference values. On the one hand, Figure S14 shows that the results from our MD simulations differ from those of Afzal et al., which clearly highlights the difficulty of merging data from different sources. On the other hand, in-depth comparisons between both sources of MD results and the experimental data, detailed in the Supporting Information, indicate good agreement between the T_g values from our MD simulations and

the experimental data. Finally, Figure 8 displays the performance of the final KNN model for the prediction of the MD-calculated T_g of the additional series of polymers. For 16 polymers the T_g is well to very well predicted, whereas the remaining 5, which are poly(1-pentene), poly(1-hexene), poly(2-heptyl acrylate), poly(vinyl propionate), and poly(N-butyl acrylamide) appear as mild outliers. The overall MAE of 23.69 K is a little higher than what was found on the testing set, but as significantly stronger outliers are visible in Figure 7, both for testing and training datasets, it is reasonable to expect the MAE to reduce over a larger number of comparisons.

4 Conclusions

This work illustrates the attractivity of combining data-driven and molecular dynamics approaches in the context of biopolymer design. As demonstrated, state-of-the art molecular dynamics simulations permit to accurately reproduce experimental trends on a key physico-chemical property of polymers, the glass transition temperature, and thanks to the recent advances in computational hardware it is possible to deploy them in a high-throughput manner. This way, the properties of 546 polymers could be evaluated *in silico* for a fraction of the time and cost of performing the equivalent experiments. In this work, 2,184 simulations were performed, for a total simulation time of 143.052 μs. Even though this number of polymers is small in comparison to the combinatorial combinations of polymeric building blocks into copolymers for the fine tuning of their properties, we have shown that it is sufficient to adequately sample the design space for a targeted application such as the replacement of the fossil-fuel based polymers used in cosmetics by biopolymers. Indeed, the ML models trained from these data show good performances and can thus drastically accelerate the exploration of the design space identified as relevant to the specific targeted application.

In that process, we have compiled a dataset constituted of 58 homopolymers, with a focus on biopolymers, and implemented a polymer building and simulation procedure permitting to

perform high-throughput simulations relying either on GLYCAM06 for polysaccharides or on GAFF for all remaining polymers. The comparison of the glass transition temperature values obtained with the simulation protocols implemented in this work and the data collected from a variety of experimental references reported in databases and the literature has evidenced that the experimental trends are well reproduced. Indeed, accounting for a well known constant offset, with a value of 105.02 K in our case, attributed to the mismatch between experimental and computational cooling rates which are orders of magnitude different, such comparison yields an R^2 score of 0.83. Then, a set of 14,877 copolymers was generated via binary combination of the collected 58 monomers with ratios in the range [0.1; 0.9] and a machine learning driven selection approach was implemented to sample the targeted design space by calculating the glass transition temperature of 546 polymers (58 homopolymers and 488 copolymers) via molecular dynamics simulations. From this dataset, a k-nearest neighbors model has been trained to predict the MD-calculated glass transition of polymers, within the design space defined by its applicability domain, with a mean absolute error of 19.34 K and an R^2 score of 0.89 when evaluated on the testing dataset isolated before the start of the training procedure.

Finally, a comparison with 14 polymers from an external dataset within the applicability domain of our ML model has confirmed that both the MD simulation protocols and the ML model are performing as good as expected. This comparison also shows the differences between two series of MD simulations with different polymer building and T_g simulation protocols, and thus highlights the difficulties of merging data from different sources. This challenge exists both when working with experimental data and with computational data and thus remains an important bottleneck for the implementation of large scale data-driven approaches in science. Therefore, it further emphasizes the great potential of mixed strategies combining data-driven and molecular dynamics approaches for targeted applications as implemented in this work. Importantly, such approaches are not limited to polymers. They can be applied to many classes of molecules or materials. Besides, the work undertaken here

could be further refined and improved by future efforts in generating additional data or in improving data quality. On the one hand, the learning curves of the ML model have shown that adding more data would permit to reduce its mean absolute error. On the other hand, even though it would require an important human effort, specifically looking at the outliers found in the predictions of both training and testing datasets could permit to increase data quality. Such outliers may be indicating some inconsistencies in a specific set of MD simulations, e.g., difficulties for the force field to handle some specific polymer chemistry, or highlighting the presence of pairs of molecules with similar structures but large difference in T_g, the so-called activity cliffs,[109] requiring a specific treatment.[110]

Acknowledgement

We acknowledge EuroHPC Joint Undertaking for awarding us access to MeluXina at LuxProvide, Luxembourg.

Conflict of Interest

The authors declare the following competing financial interest(s): A.-C. D. and F. L. are full employees of L'Oréal involved in research activities.

Supporting Information Available

Additional discussions, figures, and details of the MD simulation and ML training protocols.

References

(1) SAPEA, Science Advice for Policy by European Academies, *Biodegradability of Plastics in the Open Environment*; Berlin: SAPEA, 2020.

(2) Andrady, A. L.; Neal, M. A. Applications and societal benefits of plastics. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2009**, *364*, 1977–1984.

(3) Zheng, J.; Suh, S. Strategies to reduce the global carbon footprint of plastics. *Nat. Clim. Chang.* **2019**, *9*, 374–378.

(4) Rosenboom, J.-G.; Langer, R.; Traverso, G. Bioplastics for a circular economy. *Nat. Rev. Mater.* **2022**, *7*, 117–137.

(5) Mohanty, A. K.; Wu, F.; Mincheva, R.; Hakkarainen, M.; Raquez, J.-M.; Mielewski, D. F.; Narayan, R.; Netravali, A. N.; Misra, M. Sustainable polymers. *Nat. Rev. Methods Primers* **2022**, *2*.

(6) Nabeoka, R.; Suzuki, H.; Akasaka, Y.; Ando, N.; Yoshida, T. Evaluating the Ready Biodegradability of Biodegradable Plastics. *Environmental Toxicology and Chemistry* **2021**, *40*, 2443–2449.

(7) Kjeldsen, A.; Price, M.; Lilley, C.; Guzniczak, E.; Archer, I. *A review of standards for biodegradable plastics*; 2019.

(8) Zumstein, M. T.; Narayan, R.; Kohler, H. P. E.; McNeill, K.; Sander, M. Dos and Do Nots When Assessing the Biodegradation of Plastics. *Environmental Science and Technology* **2019**, *53*, 9967–9969.

(9) Rai, P.; Mehrotra, S.; Priya, S.; Gnansounou, E.; Sharma, S. K. Recent advances in the sustainable design and applications of biodegradable polymers. *Bioresource Technology* **2021**, *325*, 124739.

(10) Liu, J.-K. Natural products in cosmetics. *Nat. Products Bioprospect.* **2022**, *12*, 40.

(11) Patil, A.; Ferritto, M. S. *Polymers for Personal Care and Cosmetics*; Chapter 1, pp 3–11.

(12) Watkins, E.; Schweitzer, J.-P.; Leinala, E.; Börkey, P. Policy Approaches to Incentivise Sustainable Plastic Design. *OECD Environment Working Papers* **2019**, 1–61.

(13) Hayes, D.; Solaiman, D. K.; Ashby, R. D. *Biobased Surfactants*, 2nd ed.; Academic Press: San Diego, CA, 2019.

(14) Cywar, R. M.; Rorrer, N. A.; Hoyt, C. B.; Beckham, G. T.; Chen, E. Y.-X. Bio-based polymers with performance-advantaged properties. *Nat. Rev. Mater.* **2021**, *7*, 83–103.

(15) Luengo, G.; Fameau, A.-L.; Léonforte, F.; Greaves, A. *Adv. Coll. and Int. Sci.* **2021**, *290*, 102383.

(16) Coscia, B.; Shelley, J.; Browning, A.; Sanders, J.; Chaudret, R.; Rozot, R.; Léonforte, F.; Halls, M.; Luengo, G. *Phys. Chem. Chem. Phys.* **2023**, *25*, 1768.

(17) Morozova, T.; Garcia, N.; Barrat, J.-L.; Luengo, G.; Léonforte, F. *ACS Applied Materials and Interfaces* **2021**, *25*, 30086–30097.

(18) de Pablo, J.; Hillmyer, M. A. Sustainable Polymers Square Table. *Macromolecules* **2021**, *54*, 8257–8258.

(19) Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nature Reviews Materials* **2021**, *6*, 655–678.

(20) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(21) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *The Journal of Physical Chemistry C* **2018**, *122*, 17575–17585.

(22) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-learning predictions of polymer properties with Polymer Genome. *Journal of Applied Physics* **2020**, *128*, 171104.

(23) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(24) Kuenneth, C.; Schertzer, W.; Ramprasad, R. Copolymer Informatics with Multitask Deep Neural Networks. *Macromolecules* **2021**, *54*, 5957–5961.

(25) Kuenneth, C.; Lalonde, J.; Marrone, B. L.; Iverson, C. N.; Ramprasad, R.; Pilania, G. Bioplastic design using multitask deep neural networks. *Commun. Mater.* **2022**, *3*, 96.

(26) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. *Proceedings - 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011* **2011**, 22–29.

(27) Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *Journal of Chemical Information and Modeling* **2021**, *61*, 5395–5413.

(28) Ma, R.; Luo, T. PI1M: A benchmark database for polymer informatics. *Journal of Chemical Information and Modeling* **2020**, *60*, 4684–4690.

(29) NIMS Materials Database (MatNavi). `https://mits.nims.go.jp/`.

(30) Gormley, A. J.; Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nat. Rev. Mater.* **2021**, *6*, 642–644.

(31) CHEMnetBASE. `http://poly.chemnetbase.com/`.

(32) Chemical Retrieval on the Web (CROW). `https://www.polymerdatabase.com/`.

(33) Polymer Property Predictor and Database. `https://pppdb.uchicago.edu/`.

(34) Bradford, G.; Lopez, J.; Stolberg, M. A.; Osterude, R.; Jeremiah, A.; Shao-horn, Y.; Gomez-bombarelli, R. Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery. *ACS Central Science* **2022**, *9*, 206–216.

(35) Pugar, J. A.; Gang, C.; Huang, C.; Haider, K. W.; Washburn, N. R. Predicting Young's Modulus of Linear Polyurethane and Polyurethane-Polyurea Elastomers: Bridging Length Scales with Physicochemical Modeling and Machine Learning. *ACS Applied Materials and Interfaces* **2022**, *14*, 16568–16581.

(36) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun, K. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances* **2019**, *5*, 1–9.

(37) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Materials Science and Engineering R: Reports* **2021**, *144*, 100595.

(38) Martin, T. B.; Audus, D. J. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polymers Au* **2023**, *3*, 239–258.

(39) Walsh, D. J.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M. E.; Mysona, J.; Lin, T. S.; de Pablo, J. J.; Jensen, K. F.; Audus, D. J.; Olsen, B. D. Community Resource

for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. *ACS Central Science* **2023**, *9*, 330–338.

(40) Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.

(41) Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H. J.; Felser, C.; Greiner, M.; Groß, A.; Koch, C. T.; Kremer, K.; Nagel, W. E.; Scheidgen, M.; Wöll, C.; Draxl, C. FAIR data enabling new horizons for materials research. *Nature* **2022**, *604*, 635–642.

(42) Hayashi, Y.; Shiomi, J.; Morikawa, J.; Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Computational Materials* **2022**, *8*, 2–12.

(43) Musa, O. M.; Tallon, M. A. *Polymers for Personal Care and Cosmetics*; Chapter 15, pp 233–284.

(44) Gartner, T. E. I.; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* **2019**, *52*, 755–786.

(45) Patrone, P. N.; Dienstfrey, A.; Browning, A. R.; Tucker, S.; Christensen, S. Uncertainty quantification in molecular dynamics studies of the glass transition temperature. *Polymer* **2016**, *87*, 246–259.

(46) Buchholz, J.; Paul, W.; Varnik, F.; Binder, K. Cooling rate dependence of the glass transition temperature of polymer melts: Molecular dynamics study. *Journal of Chemical Physics* **2002**, *117*, 7364–7372.

(47) Mohammadi, M.; Fazli, H.; Karevan, M.; Davoodi, J. The glass transition temperature of PMMA: A molecular dynamics study and comparison of various determination methods. *European Polymer Journal* **2017**, *91*, 121–133.

(48) Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer informatics with multi-task learning. *Patterns* **2021**, *2*, 100238.

(49) Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns (N. Y.)* **2021**, *2*, 100225.

(50) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *Journal of Chemical Information and Modeling* **2019**, *59*, 5013–5025.

(51) Afzal, M. A. F.; Browning, A. R.; Goldberg, A.; Halls, M. D.; Gavartin, J. L.; Morisato, T.; Hughes, T. F.; Giesen, D. J.; Goose, J. E. High-Throughput Molecular Dynamics Simulations and Validation of Thermophysical Properties of Polymers for Various Applications. *ACS Applied Polymer Materials* **2021**, *3*, 620–630.

(52) Bejagam, K. K.; Iverson, C. N.; Marrone, B. L.; Pilania, G. Molecular dynamics simulations for glass transition temperature predictions of polyhydroxyalkanoate biopolymers. *Phys. Chem. Chem. Phys.* **2020**, *22*, 17880–17889.

(53) Vanommeslaeghe, K.; MacKerell, A. D. J. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *Journal of Chemical Information and Modeling* **2012**, *52*, 3144–3154.

(54) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. J. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *Journal of Chemical Information and Modeling* **2012**, *52*, 3155–3168.

(55) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(56) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260.

(57) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 367.

(58) Alesadi, A.; Cao, Z.; Li, Z.; Zhang, S.; Zhao, H.; Gu, X.; Xia, W. Machine learning prediction of glass transition temperature of conjugated polymers from chemical structure. *Cell Reports Physical Science* **2022**, *3*, 100911.

(59) Andrews, J.; Handler, R. A.; Blaisten-Barojas, E. Structure, energetics and thermodynamics of PLGA condensed phases from Molecular Dynamics. *Polymer* **2020**, *206*, 122903.

(60) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry* **2008**, *29*, 622–655.

(61) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **1984**, *106*, 765–784.

(62) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **1986**, *7*, 230–252.

(63) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, B. Molecular Mechanical and Molecular Dynamic Simulations of Glycoproteins and Oligosaccharides. 1. GLYCAM_93 Parameter Development. *The Journal of Physical Chemistry* **1995**, *99*, 3832–3846.

(64) Banerjee, P.; Silva, D. V.; Lipowsky, R.; Santer, M. Editor ' s Choice The importance of side branches of glycosylphosphatidylinositol anchors : a molecular. *Glycobiology* **2022**, *32*, 933–948.

(65) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.

(66) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry* **2000**, *21*, 132–146.

(67) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

(68) Ratcliff, L. E.; Mohr, S.; Huhs, G.; Deutsch, T.; Masella, M.; Genovese, L. Challenges in large scale quantum mechanical calculations. *WIREs Computational Molecular Science* **2017**, *7*, e1290.

(69) Fortunato, M. E.; Colina, C. M. pysimm: A python package for simulation of molecular systems. *SoftwareX* **2017**, *6*, 7–12.

(70) Demidov, A. G.; Perera, B. L. A.; Fortunato, M. E.; Lin, S.; Colina, C. M. Update 1.1 to "pysimm: A python package for simulation of molecular systems", (PII: S2352711016300395). *SoftwareX* **2021**, *15*, 100749.

(71) Santana-Bonilla, A.; López-Ríos de Castro, R.; Sun, P.; Ziolek, R. M.; Lorenz, C. D. Modular Software for Generating and Modelling Diverse Polymer Databases. *Journal of Chemical Information and Modeling* **2023**, *63*, 3761–3771.

(72) Degiacomi, M. T.; Erastova, V.; Wilson, M. R. Easy creation of polymeric systems for molecular dynamics with Assemble! *Computer Physics Communications* **2016**, *202*, 304–309.

(73) Ramos, M. C.; Quoika, P. K.; Horta, V. A. C.; Dias, D. M.; Costa, E. G.; do Amaral, J. L. M.; Ribeiro, L. M.; Liedl, K. R.; Horta, B. A. C. pyPolyBuilder: Automated Preparation of Molecular Topologies and Initial Configurations for Molecular Dynamics Simulations of Arbitrary Supramolecules. *Journal of Chemical Information and Modeling* **2021**, *61*, 1539–1544.

(74) QuantumATK Polymer Builder. `https://docs.quantumatk.com/tutorials/polymer_builder/polymer_builder.html`.

(75) Schrödinger Polymeric Materials. `https://www.schrodinger.com/science-articles/polymeric-materials`.

(76) Grünewald, F.; Alessandri, R.; Kroon, P. C.; Monticelli, L.; Souza, P. C. T.; Marrink, S. J. Polyply; a python suite for facilitating simulations of macromolecules and nanomaterials. *Nat. Commun.* **2022**, *13*, 68.

(77) Horta, B. A.; Merz, P. T.; Fuchs, P. F.; Dolenc, J.; Riniker, S.; Hünenberger, P. H. A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set. *Journal of Chemical Theory and Computation* **2016**, *12*, 3825–3850.

(78) Abbott, L. J.; Hart, K. E.; Colina, C. M. Polymatic: A generalized simulated polymerization algorithm for amorphous polymers. *Theoretical Chemistry Accounts* **2013**, *132*, 1–19.

(79) Sahu, H.; Shen, K.-H.; Montoya, J. H.; Tran, H.; Ramprasad, R. Polymer Structure Predictor (PSP): A Python Toolkit for Predicting Atomic-Level Structural Models for

a Range of Polymer Geometries. *Journal of Chemical Theory and Computation* **2022**, *18*, 2737–2748.

(80) BIOVIA Material Studio. `https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/biovia-materials-studio/polymer-composites/`.

(81) Rappe, A. K.; Goddard III, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358–3363.

(82) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.

(83) Grest, G. S.; Lacasse, M.-D.; Kremer, K.; Gupta, A. M. Efficient continuum model for simulating polymer blends and copolymers. *J. Chem. Phys.* **1996**, *105*, 10583–10594.

(84) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*, 108171.

(85) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A. I.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(86) Halgren, T. A. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17*, 490–519.

(87) Larsen, G. S.; Lin, P.; Hart, K. E.; Colina, C. M. Molecular Simulations of PIM-1-like Polymers of Intrinsic Microporosity. *Macromolecules* **2011**, *44*, 6944–6951.

(88) Hofmann, D.; Fritz, L.; Ulbrich, J.; Schepers, C.; Böhning, M. Detailed-atomistic

molecular modeling of small molecule diffusion and solution processes in polymeric membrane materials. *Macromolecular Theory and Simulations* **2000**, *9*, 293–327.

(89) Karayiannis, N. C.; Mavrantzas, V. G.; Theodorou, D. N. Detailed Atomistic Simulation of the Segmental Dynamics and Barrier Properties of Amorphous Poly(ethylene terephthalate) and Poly(ethylene isophthalate). *Macromolecules* **2004**, *37*, 2978–2995.

(90) Bandyopadhyay, S.; Anil, A. G.; James, A.; Patra, A. Multifunctional Porous Organic Polymers: Tuning of Porosity, CO2, and H2 Storage and Visible-Light-Driven Photocatalysis. *ACS Applied Materials & Interfaces* **2016**, *8*, 27669–27678.

(91) Li, C.; Medvedev, G. A.; Lee, E.-W.; Kim, J.; Caruthers, J. M.; Strachan, A. Molecular dynamics simulations and experimental studies of the thermomechanical response of an epoxy thermoset polymer. *Polymer* **2012**, *53*, 4222–4230.

(92) Cologne, J.; Sposto, R. Smooth piecewise linear regression splines with hyperbolic covariates. *Journal of Applied Statistics* **1994**, *21*, 221–233.

(93) Werpy, T.; Petersen, G. *Top Value Added Chemicals from Biomass - Volume I: Results of Screening for Potential Candidates from Sugars and Synthesis Gas*; 2004.

(94) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4.

(95) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**,

(96) Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(97) Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition

temperatures. *Modelling and Simulation in Materials Science and Engineering* **2019**, *27*, 24002.

(98) Fox, T. G. Influence of diluent and of copolymer composition on the glass temperature of a polymer system. *Bull. Am. Phys. Soc.* **1952**, *1*, 123.

(99) Huang, C. C.; Du, M. X.; Zhang, B. Q.; Liu, C. Y. Glass Transition Temperatures of Copolymers: Molecular Origins of Deviation from the Linear Relation. *Macromolecules* **2022**, *55*, 3189–3200.

(100) Drayer, W. F.; Simmons, D. S. Sequence Effects on the Glass Transition of a Model Copolymer System. *Macromolecules* **2022**, *55*, 5926–5937.

(101) Molnar, C. *Book*, leanpub ed.; 2020; p 247.

(102) OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; 2014; p 154.

(103) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791–4810.

(104) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSAR Models. *International Journal of Quantitative Structure-Property Relationships* **2016**, *1*, 45–63.

(105) He, Z.; Xu, X.; Deng, S. Discovering cluster-based local outliers. *Pattern Recognition Letters* **2003**, *24*, 1641–1650.

(106) Breunig, M. M.; Kriegel, H. P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* **2000**, 93–104.

(107) Zhao, Y.; Nasrullah, Z.; Li, Z. PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research* **2019**, *20*, 1–7.

(108) Bicerano, J. In *Prediction of Polymer Properties*, third edition ed.; Inc, M. D., Ed.; 2002; p 746.

(109) Van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *Journal of Chemical Information and Modeling* **2022**, *62*, 5938–5951.

(110) Dablander, M.; Hanser, T.; Lambiotte, R.; Morris, G. M. Exploring QSAR Models for Activity-Cliff Prediction. *Journal of Cheminformatics* **2023**, *15*, 47.

TOC Graphic