Ml-Gle: A Machine Learning Enhanced Generalized Langevin Equation Framework for Transient Anomalous Diffusion in Polymer Dynamics

Gian-Michele Cherchi, Alain Dequidt, Vincent Barra, Arnaud Guillin, Patrice Hauret, Nicolas Martzel

Journal of Computational Physics (2024)

ML-GLE: a machine learning enhanced Generalized Langevin equation framework for transient anomalous diffusion in polymer dynamics

Gian-Michele Cherchi[a], Alain Dequidt[a], Arnaud Guillin[b], Nicolas Martzel[c], Patrice Hauret[c], Vincent Barra[d]

[a]*Université Clermont-Auvergne, CNRS, ICCF, F-63000 Clermont–Ferrand, France.*
[b]*Université Clermont-Auvergne, CNRS, LMBP, F-63000 Clermont–Ferrand, France.*
[c]*Manufacture Française des Pneumatiques Michelin, 63040 Clermont-Ferrand, France.*
[d]*Université Clermont-Auvergne, Clermont Auvergne INP, CNRS, Mines de Saint-Étienne, LIMOS, Clermont-Ferrand F-63000, France*

ARTICLE INFO

Article history:

Scientific machine learning
Molecular dynamics
Autoregressive models
Coarse-graining
Neural network
Stochastic modeling

ABSTRACT

In this work we introduce ML-GLE, a machine learning framework to generate long-term single-polymer dynamics by exploiting short-term trajectories from molecular dynamics (MD) simulations of polymer melts. Even with current advances in machine learning for MD, these polymeric materials remain difficult to simulate and characterize due to prohibitive computational costs when long relaxation timescales are involved. Our method relies on a 3D neural auto-regressive model for single polymer lower dimensional collective variables, called normal modes. This enhances the Generalized Langevin Equation (GLE) capabilities in modeling diffusion phenomena. We exploit a particular GLE solution which is known to reproduce the mean square displacement curve relative to transient anomalous diffusion and connect it with the normal modes collective variables. ML-GLE is capable of emulating the single polymer statistical properties in the long-term, predicting the diffusion coefficient. As a consequence, this results in an enormous acceleration in terms of simulation time with respect to the full-size simulation. Moreover, this approach is also scalable with system size.

1. Introduction

Multiscale modeling has emerged as an invaluable approach for simulating systems with a great number of degrees of freedom (d.o.fs) and whose components are organized in a hierarchical structure (e.g. proteins, polymers, fluids). These systems display important properties at multiple time/space scales. Consequently, the phenomena of interest are often out-of-reach because they are the result of collective dynamics and complex non-linear interactions that require a considerable computational cost for a detailed *fine-grained* simulation (e.g. atomistic), resolving processes at a short time-scale, for which the governing laws are well known.
Consequently, it becomes difficult to assess structural and dynamical properties of the system under study. Multiscale techniques focus therefore on developing multiple models that can be used either concurrently to describe the system at different scales, or sequentially, profiting from simulations at finer scales to inform a *coarse-grained* (CG) at larger scales, with less d.o.fs.

In this respect, machine learning (ML) and neural networks (NNs) have proved useful in bridging the gap between different spatial and temporal timescales. If the system can be described with Partial Differential Equations (PDEs), physics informed neural networks (PINNs) can be employed [1, 2], by exploiting *a priori* knowledge about the governing laws of the physical system together with physical principles in order to approximate the unknown solution of a (possibly) high-dimensional PDE. The PDE solution is constructed by means of a NN, which is trained on data

e-mail: alain.dequidt@uca.fr (Alain Dequidt)

coming from either real world measurements or simulation data in combination with the boundary conditions (BCs). The information about the governing laws is provided to the learning framework through a regularizer, guiding the NN solution towards an optimal high fidelity approximation. Furthermore, the expressiveness of NNs favors the solution's generalization capability, allowing for the PINNs to be integrated at different spatial resolutions without retraining, unlocking therefore longer timescales with useful applications in forecasting and control [3, 4]. PINNs can be also combined with data-driven reduced-order models (ROMs) [5] in order to evaluate the PDE parametric dependence efficiently using surrogates while at the same time surmounting data scarcity issues.

More often, the laws governing the dynamical system are unknown and the inverse problem, inferring the model from observations or simulated time series is a much challenging one. Given the large number of d.o.fs involved, dimensional reductions are often needed to make the system treatable.

The dynamic mode decomposition (DMD) technique [6] simultaneously extracts meaningful hierarchical dynamical features from multidimensional time series data and approximates their time evolution through a linear operator, **P**. In case of linear systems, these features reduce to the normal modes of fluctuation. More complicated methods are able to identify the most parsimonious model governing the non-linear dynamics [7], and also approximate PDEs, from Navier-Stokes to Brownian motion [8, 9].

In the context of molecular dynamics (MD), dimensional reduction techniques are based on the projection operator [10], for a proper selection of relevant CG variables. Data-driven techniques exist for efficient and transferable molecular CG variables, employing NNs [11, 12, 13, 14, 15], but they are limited to small molecular or condensed matter systems, which in turn allow for the *Markov* approximation to be reasonable, disregarding in this way any memory effect.

The latter can be included through the Mori-Zwanzig (MZ) formalism [16, 17, 18, 19], which takes advantage of the projection operator to obtain dynamical equations for the *non-Markovian* dynamics of the CG variables. Thus, MZ allows to derive closed form expressions for the CG variables by treating the irrelevant d.o.fs as stochastic noise while the unresolved variables are accounted for with a time-dependent and in general unknown memory kernel function. In practice, identifying meaningful lower dimensional collective variables (CVs) and ensuring CG dynamics to be consistent at different thermodynamic conditions is an arduous task and system-dependent. With MZ as a starting point, further *ansatzes* are then required, where the usual procedure consists in deriving the Generalized Langevin Equation (GLE) [20], for some CVs.

For some systems, even good quality CG variables are not enough to observe and measure the quantities of interest. In case of visco-elastic systems like polymer melts (Fig. 1.left), the strong non-linear interactions between polymers and the topological constraints result in an extremely slow polymer dynamics.

Single polymers undergo *Transient Anomalous Diffusion* (TAD) [21, 22, 23, 24, 25], which can be identified by computing the mean square displacement (MSD). Given a generic spatial variable x, one has, $\langle x^2(t)\rangle \propto t^{\nu}$ at short time-scales, while asymptotically $\langle x^2(t)\rangle \propto t$, up to a proportionality constant D, called diffusion coefficient.

The two regimes describe the transient anomalous part, and the normal Brownian diffusive one. If $\nu \in (0, 1)$ and $\nu \in (1, +\infty)$, one has respectively sub-diffusive and super-diffusive behaviour. The transient period can be long before observing normal diffusion. An example is shown in Fig 1.Right.

Resorting to the description of single polymer chains and their *effective* dynamics in the melt reduces dramatically the number of d.o.fs, and the governing dynamics can be described by a GLE. However, the associated *non-Markovian* character of GLEs makes memory effects difficult to model and reproduce faithfully. Parametric models of TAD employing GLEs exist but they fail in the small data regime and consequently no information is available on the transient and asymptotic regime [26, 27]. In these cases, data availability jeopardise any possibility of applying data-driven techniques, because the underlying processes characterizing the long-term behaviour of the system cannot be observed. Estimating diffusion-related properties seems therefore unfeasible. Several studies have proposed a data-driven parametrization of GLEs [28, 29, 30, 31, 32, 33, 34, 35, 36, 37] but they typically directly deal with the memory kernel additionally equipping the description with auxiliary variables characterizing an extended Markovian dynamics. Nonetheless, applications are restricted to small molecules and are unable to reconstruct the original fine-grained system. Another similar data-driven approach consists in modeling long-time *non-Markovian* dynamics through an integrative generalized master equation (IGME) [38], utilizing directly the time integral of memory kernel through Taylor series expansion to model conformational transitions in small complexed biological molecules.

Chorin and Lu's seminal work [39] has outlined the connection between nonlinear auto-regressive models and the governing equations related to the stochastic dynamics of a subset of variables of interest, connecting the latter description to the MZ formalism in statistical physics. In [40], a GLE description based on a nonlinear auto-regressive

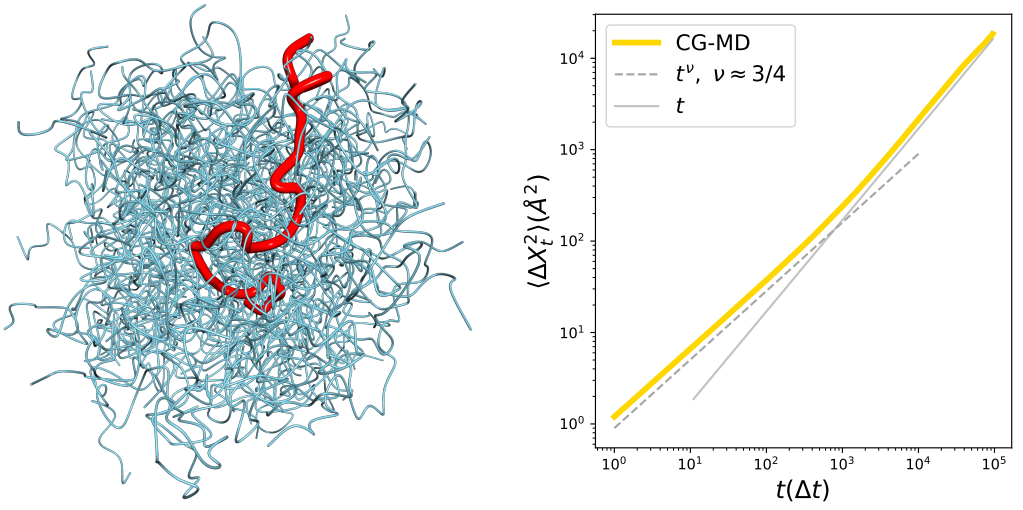

Fig. 1: (**Left**) Rendering of an homo-polymer melt *Butadiene Rubber* simulation snapshot, used as benchmark system for the ML-GLE. The reference simulations are done using the monomers as CG variables integrated with a small timestep δt, as explained in sec. 4.1. In red, a generic **tagged** polymer, whose dynamics is instead extrapolated with a bigger timestep $\Delta t \gg \delta t$. (**Right**) Example of an MSD curve relative to single polymer Center of Mass, averaged over all polymer chains, in log-log scale. The long transient regime is clearly visible, before reaching a normal diffusive dynamics in the long-time.

model for a 1D CV is proposed, with data coming from an *ab initio* MD simulation of a magnetic system.

Along these lines, this work proposes, to our knowledge, the first application of a data-driven GLEs employing NNs, to a realistic 3D diffusion phenomena in soft matter. We present an *hybrid* ML-GLE framework for simulating the effective dynamics of single polymer d.o.fs inside melt and a concrete case in which the GLE can be approximated in a data-driven way. With minimal assumptions on the kernel memory function, the GLE solution can be used to perform both long-term stable single polymer simulations and make predictions about the transient and asymptotic diffusive regime without running full-size simulations, which would require several orders of magnitude more simulation time-steps.

Our framework accelerates single polymer stochastic dynamics by decomposing its d.o.fs in a set of slow CVs: the center of mass (C.o.M) diffusive dynamics is modeled with a parameterized solution of an *ansatz* GLE, which is known for reproducing TAD [27], while faster physically motivated 3D CVs, called normal modes, are accounted for by a neural auto-regressive (NAR) model, able to approximate the *coarse-grained* time integrator of *downsampled* MD trajectories from an already CG polymer melt simulation. In addition, by exploiting the symmetries of the physical system, the 3D CV stochastic dynamics is stable in the long term, since the NAR model implicitly endows the rotational symmetries naturally present in the physical system. Training and generation are fast, allowing the computational cost to estimate asymptotic relaxation properties (MSD) to be reduced dramatically.

Related work on extrapolating long term behaviour with NNs from MD trajectories has been done in the past; in [41], the authors propose a Generative Adversarial Network (GAN) for sequence to sequence generation with data coming from different MD simulations, including polymer melts. The method appears to be successful in reproducing TAD in one system, but there is no net gain in terms of computational cost. In addition, GANs are notoriously difficult to train and interpret. In our case, since the GLE description exploits normal modes of fluctuation, interpretability is assured. In [42], the authors propose to learn *non-Markovian* stochastic processes with recurrent NNs by optimizing on non-parametric statistical estimators, like the kernel density estimator (KDE) for the stationary PDFs. This method is efficient in extrapolating long-term behaviour, but *a priori* bandwidth selection for KDE introduces additional free parameters to be optimized, especially when dealing with multivariate data. By contrast, our approach attempts to directly approximate the process conditional distribution by using a maximum-likelihood parametric approach based on NNs, without any prior assumption on the long-term statistical properties.

The paper is organized in the following way. In section 2 we review the conditions under which the GLE is capable of reproducing TAD, and provide an analytical solution, depending on some Markovian CVs. Inspired by the latter, in section 3 we propose the ML-GLE framework for single polymer dynamics, which combines a parametric extended

3D GLE solution with a set of independent NAR models for the *non-Markovian* CVs, which we identify with the single polymer normal modes. In section 4 we validate the framework by training modes data on MD trajectories of an homo-polymer melt at different temperatures. The validation of the GLE parametrization is shown in section 4.5. We demonstrate the model's predictive capabilities in section 4.4 when generating long-time mode dynamics while in section 4.6 we prove the that ML-GLE reproduces correctly TAD from short-time MD data, with an effective gain in computational time with respect to the reference full-size simulation.

2. Transient anomalous diffusion and the Generalized Langevin equation

The Generalized Langevin Equation (GLE) is a stochastic integral differential equation first derived in the context of transport, as an extension of the more popular Langevin equation describing the Brownian motion of a particle in an implicit fluid [10, 43]. The GLE became a popular framework for modeling diffusion through a viscoelastic medium, where the dissipative term is not instantaneous but has a time-lagged effect, resulting in a *non-Markovian* dynamics and where the environment is described implicitly with a memory term.
In [20, 16] the same equation was derived from an arbitrary Hamiltonian non-linear interacting system, exploiting the Projection Operator formalism, prior developed by Mori [10]. The MZ formalism performs therefore a dimensional reduction, projecting the dynamics of the original system onto a selection of CVs.
Starting usually from microscopic d.o.fs and defining the CVs of interest, their *non-Markovian* effective dynamics can be expressed as a complicated stochastic differential equation, with a kernel function accounting for memory effects. In this section, we show how, given some appropriate assumptions, the GLE can account for TAD and an analytical solution can obtained in the zero-mass limit. The solution can be then fitted exploiting short time C.o.M trajectories coming from a MD simulation.

Let $\mathbf{X}(t)$ be a generic 3D CV of a particle undergoing TAD (e.g. C.o.M). The *ansatz* GLE reads as follows,

$$M\frac{d^2\mathbf{X}(t)}{dt^2} = \mathbf{K}(\mathbf{X}(t),t) - \int_0^t \mathbf{\Gamma}(t-s)\frac{d\mathbf{X}(s)}{dt}ds + \mathbf{F}(t) \tag{1}$$

A few assumptions are necessary at this point:

1. $\mathbf{\Gamma} : \mathbb{R} \to \mathbb{R}^{3\times3}_{+}$ is an homogeneous and time independent kernel function. As the environment acts as a thermal bath for the particle, we assume the fluctuation-dissipation theorem (FDT) holds, $\mathbb{E}[\mathbf{F}(t)\mathbf{F}^T(s)] = k_B T\mathbf{\Gamma}(|t-s|)$, for a fixed temperature T.
2. $\mathbf{F}(t)$ is a 3D colored noise such that $\mathbb{E}[\mathbf{F}(t)] = 0$ and $\mathbb{E}[\mathbf{F}(t)(d\mathbf{X}(0)/dt)^T] = 0$. The former implies an *isotropic* environment while the latter expresses the absence of linear correlations with respect to the initial conditions.
3. $\mathbf{K} : \mathbb{R}^3 \times [0,\infty) \to \mathbb{R}^{3\times3}$ is an external driving force and/or a conservative force. For the purposes of this work, since we are treating diffusion at equilibrium, this term is identically zero.

The information about the environment is implicitly contained in the memory kernel, the form of which is seldom known and in general intractable. Nevertheless, the functional properties of the latter determine the behaviour of the MSD in the integrated solution. This connection was initially studied by Morgado et al. [44] and further investigated in detail by McKinley and Nguyen [45], where the authors prove an important implication between the kernel properties and the asymptotic behaviour of the MSD.
Let $\{X(t)\}_{t\geq0}$ be the 1D GLE solution. The *meta-theorem* states that if the kernel function $\Gamma(t)$ is *integrable*[1], then $\lim_{t\to\infty}\langle X^2(t)\rangle = 2Dt$, where D is the diffusion coefficient. This theorem characterizes the features a kernel function must possess so that the GLE reproduces TAD. A detail proof can be found in [45]
Broadly speaking, if the kernel function is *integrable* one observes an asymptotic normal diffusive behaviour (Fig. 1.right). In practice, the transient regime can be very long before witnessing normal diffusion.
A process which is asymptotically diffusive can display one or more prior anomalous diffusion regimes: in case of polymer melts for example, one can identify up to three sub-diffusive regimes in the monomer and C.o.M dynamics, depending on different factors (for details see [46]).

[1] More details on kernel families and integrability conditions, as well as TAD can be found in [45]

2.1. Solution in the zero-mass limit with an integrable memory kernel

In this section, for simplicity, we provide a solution for the 1D version of the *ansatz* GLE presented in Eq. (1), in the absence of external forces ($\mathbf{K}(\mathbf{X}, t) = 0$). The first step consists in specifying the functional form of the memory function.

Following the results of the *meta-theorem*, in order to reproduce TAD, $\Gamma(t)$ must be an integrable function. As a result, a good candidate for the kernel *ansatz* can be the popular *Prony Series*, as proposed in [27],

$$\Gamma_N(t) = \eta \sum_{j=1}^{N} e^{-\gamma_j t}, \tag{2}$$

where $\{\gamma_j\}_{j=1}^N$ is the diffusive spectrum sequence and uniform weight η . As we will see in sec.3, this is well justified for polymer systems, and has a clear physical interpretation when dealing with classical polymer models, like the Rouse or Zimm model.

It is known that a sum of exponential functions can asymptotically approximate any *power-law* behaviour and a different one near the origin [27, 47], with sufficiently many terms. In addition, Eq. (2) belongs to the class of *complete monotone functions* which are always *integrable*. As a consequence, it can be shown that the associated GLE solution always displays asymptotic diffusive MSD [45]. Furthermore, kernel convexity $\Gamma(t) \in C^2(0, \infty)$ and *monotonicity* for $\Gamma''(t)$ near the origin are required in the zero-mass limit.

An analytic solution to the GLE can be obtained by specifying the noise term $F(t)$ [27]. Decomposing the latter as a sum of Ornstein-Uhlenbeck (OU) processes $\{F_j\}_{j=1}^N$, one for each diffusive timescale of the Prony series, the FDT takes the form, $\forall (i, j) \in \{1, ..., N\}^2$,

$$\mathbb{E}[F_i(t)F_j(s)] = \delta_{ij} k_B T \eta e^{-\gamma_j |t-s|}, \tag{3}$$

where δ_{ij} is the Kronecker delta. $F_j(t)$ can be expressed with its formal stochastic integral solution,

$$F_j(t) = \sqrt{2k_B T \eta \gamma_j} \int_0^t e^{-\gamma_j (t-s)} dW_j(s), \tag{4}$$

where W_j are independent Brownian motions.

Given a generic continuous function, $g(t) : \mathbb{R}_+ \to \mathbb{R}^3$, we define its unilateral *Laplace* transform as,

$$\mathcal{L}\{g\}(z) = \tilde{g}(z) = \int_0^\infty g(t) e^{-zt} dt \tag{5}$$

Applying the definition to Eq. (1) in 1D, one can solve for position in the Laplace domain, yielding,

$$\tilde{X}(z) = z^{-1}(Mz + \tilde{\Gamma}(z))^{-1} \tilde{F}(z) \tag{6}$$

Denoting with $\chi(t) = \mathcal{L}^{-1}\{(Mz + \tilde{\Gamma}(z))^{-1}\}$ and $G(t) = \mathcal{L}^{-1}\{z^{-1}\tilde{F}(z)\}$, and assuming vanishing initial position and velocity in time domain, the GLE solution reads,

$$X(t) = \int_0^t \chi(t-s) G(s) ds \tag{7}$$

It is important to highlight the importance of the uniform Prony series assumption from a modeling perspective. In the zero-mass limit, this choice allows for the analytical inversion of the Laplace transform back to time domain. A more detailed derivation is provided in Appendix A while we refer to [27] for additional information.

Further denoting with $\sigma^2 = 2\eta^{-1} k_B T$ and $\bar{\tau} = \frac{1}{N} \sum_{j=1}^N \gamma_j^{-1}$, the solution reads as follows,

$$X_N(t) = \frac{\sigma}{\sqrt{N\bar{\tau}}} B(t) + \sigma \sum_{j=1}^{N-1} c_j z_j(t), \tag{8}$$

where $B(t)$ is a standard Brownian motion accounting for the slower translational diffusion through the pre-factor $1/\sqrt{\bar{\tau}}$, while $\{z_j\}_{j=1}^{N-1}$ and $\{c_j\}_{j=1}^{N-1}$ are respectively a set of increasingly faster and independent OU processes and coefficients both derived from the whole diffusive spectrum. These processes are very important as they implicitly carry

information about the environment but may also describe collective dynamics pertaining to the system under study, since they are obtained directly from $\{\gamma_j\}_{j=1}^N$.
In generic complex systems, identifying these "hidden" d.o.fs and accessing their dynamics from simulations or experiments is a challenging task; in the following we will see that for visco-elastic systems such as polymer melts, these d.o.fs correspond to single polymer CVs and no additional environmental information is needed.

Parameter inference on Eq. (8) can be performed with only three parameters, by specifying the diffusive scaling as a power-law, called the Generalized Rouse Kernel (GRK), and introduced in [48]. One has, $\forall\, j \in \{1, ..., N\}$,

$$\gamma_j = \left(\frac{j}{N}\right)^{\rho} \tau_1^{-1}, \tag{9}$$

where τ_1 is the shortest relaxation time-scale relative to and $\rho \in (0, \infty)$.
The MSD can be derived from Eq. (8), exploiting the Rouse kernel and the properties of the OU process, yielding,

$$\mathbb{E}[X_N^2(t)] = \frac{\sigma^2}{N\bar{\tau}} t + \tau_1 \sum_{j=1}^{N} \left(\frac{N}{j}\right)^{\rho} \left(1 - e^{-(j/N)^{\rho} t/\tau_1}\right) \tag{10}$$

where we took expectations over the path space.
It can be shown that Eq. (10) is capable of reproducing any TAD exponent with a single anomalous regime, whose characteristics are uniquely determined by the parameters ρ, $\bar{\tau}$ and τ_1.
The power-law exponent ρ is related to the anomalous exponent ν by the following relation,

$$\nu = 1 - \frac{1}{\rho} \tag{11}$$

which can be derived from the sum on the right hand side of Eq. (10) for sufficiently large N (see [27]).
In practice, Eq. (8) displays a remarkable robustness to reproduce the same C.o.M scaling behaviour when random perturbations are applied to the coefficients $\{c_j\}_{j=1}^{N-1}$. As demonstrated by McKinley et al. [27], denoting the coefficients uniform second moment as $\sigma_c^2 = \mathbb{E}[c_k^2]$, one has that,

$$Var(\mathbb{E}[X_N^2(t)]) \sim \frac{\sigma_c^2}{N} + o\left(\frac{1}{N}\right) \tag{12}$$

It follows from this argument that, when conducting inference from real or simulated data, as long as the coefficients are statistically independent and of the same magnitude, they do not impact the intermediate time-scale and its corresponding anomalous exponent, which is only determined the modes correlation. This will also impact our modeling choices as shown in the next section. However, inference with this model requires data at long timescales needed to fit the solution in the anomalous and asymptotic regime. Nonetheless, even in the simple case of polymer dynamics from which it is inspired, it possesses a few limitations.

1. It restricts the family of completely monotone functions to a particular functional form through the power-law scaling; in reality, the memory kernel can have an arbitrary non-trivial expression, with both a weight profile $\{\eta_j\}_{j=1}^N$ and a diffusive spectrum $\{\gamma_j\}_{j=1}^N$ depending on many factors, like temperature, density, chain length or chemical species.
2. Application is limited to 1D data sources and prediction of transient and asymptotic properties on real 3D physical systems, from experiments or simulations, remains unattainable without running full-size simulations or having access to long-time trajectories.

In this work, we apply the GLE framework on CG MD simulation data, modeling single polymer C.o.M dynamics inside the polymer melt, by extending Eq. (8) to the 3D case. We further employ a set of independent NAR models for collective single polymer d.o.fs (normal modes), and show how they can both accelerate single polymer d.o.fs simulation reproducing memory effects and estimate C.o.M diffusion properties by using only a fraction of simulation time. As a result, no assumptions are made on the kernel function but the integrability condition, since the whole procedure is data-driven. Nevertheless, the proposed method is able to reproduce the MSD curve from short MD trajectories and simulate long-time dynamics.

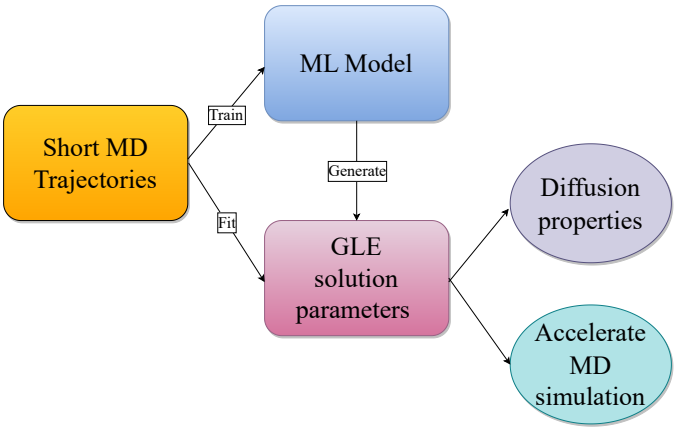

Fig. 2: Block diagram of the ML-GLE framework. Short MD trajectories are used both to train the NAR generative model and to fit the solution to the GLE equation. This allows to simulate an effective single polymer dynamics as if it came from the polymer melt simulation, as well as assess its long-term behaviour. The physics informed GLE solution connects different CVs, allowing the acceleration of single polymer dynamics.

3. ML-GLE Framework

The developed framework ML-GLE is shown in Fig. 2. It generates single polymers dynamics emulating its behaviour inside a polymer melt. It is a data-driven *hybrid* method as it exploits the solution of an *ansatz* GLE, able to reproduce TAD, to model the C.o.M. The "physics informed" GLE solution connects the C.o.M dynamics to a set of faster single polymer CVs. Their conditional distribution is approximated with independent 3D NAR generative models, trained on short-time MD trajectories, each of them accounting for a different 3D CV.
It is easy to extend Eq. (8) to the 3D case for the C.o.M dynamics. By assuming a uniform pre-factor $c \forall j \in \{1, ..., N-1\}$, one can define $\beta = \sigma c$, hence obtaining,

$$\mathbf{X}_k(t) = \alpha \mathbf{B}(t) + \beta \sum_{j=1}^{k} \mathbf{z}_j(t), \tag{13}$$

where $\alpha, \beta \in \mathbb{R}_+$ are now two parameters, $\mathbf{B}(t) \in \mathbb{R}^{3\times 1}$ is a 3D standard Brownian motion, and $\{\mathbf{z}_j(t)\}_{j=1}^{k} \in \mathbb{R}^{3\times 1}$ is the set of 3D CVs stochastic processes (*non-Markovian*). As shown in the previous section, setting a uniform coefficient β does not compromise the MSD scaling behaviour and also allows us to have a more parsimonious model instead of fitting a set of heterogeneous coefficients.
In the prototypical case of polymer dynamics, one can notice there exists a similarity between the latter and single polymer modes of fluctuation [46], called *normal modes*. In fact, in the classical theory of self-intersecting *phantom* chains, the Rouse model describes chain dynamics by decomposing its configuration in a set of *Markovian* OU processes $\{\mathbf{z}_j\}_{j=0}^{N-1}$, called *Rouse modes*, exemplified in Fig. 3.
For every monomer variable $\{\mathbf{x}_n\}_{n=1}^{N}$, the Rouse model prescribes the following decomposition,

$$\mathbf{x}_n = \mathbf{z}_0 + 2 \sum_{j=1}^{N-1} \mathbf{z}_j \cos\left[\frac{j\pi}{N}\left(n + \frac{1}{2}\right)\right], \tag{14}$$

where $\mathbf{z}_0$ is the C.o.M variable. The structure of Eq. (14) is reminiscent of the same decomposition in Eq. (13). Furthermore, in the Rouse model the OU processes model indeed the modes dynamics reproducing the monomer anomalous exponent 1/2. The choice of uniform weights kernel function can also be understood in terms of Rouse modes, whose correlations are exponential parameterized by the diffusive scaling $\gamma_j \sim \sin^2(j\pi/2N), \forall j \in \{1, ..., N-1\}$ (see [27, 46]). It seems natural therefore that the *non-Markovian* normal modes dynamics obtained from a chain

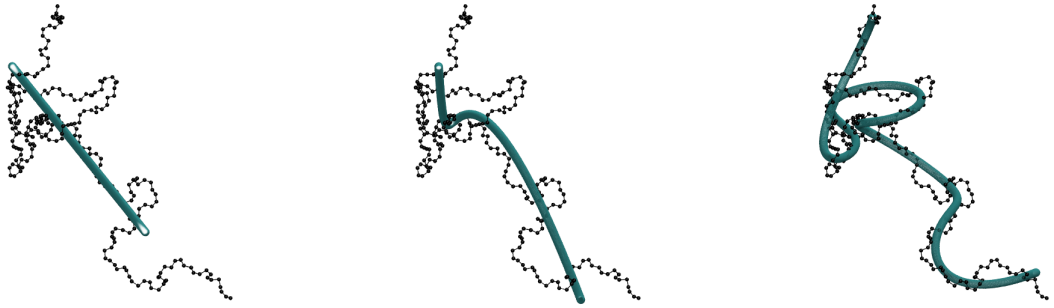

Fig. 3: Depiction of a phantom chain configuration simulated with the Rouse model, and its respective reconstructed profile with $k = 2$, $k = 4$ and $k = 12$ modes (from Left to Right). Notice how the interpolation becomes more accurate while increasing the number of slower modes k, used for reconstruction. Their role pertains to collective polymer motion and their *non-Markovian* dynamics is essential to capture asymptotic behaviour.

configuration in a realistic simulation would be relevant for reproducing the C.o.M diffusive dynamics.
Thus, we assume that sufficient information about the long-term behaviour is contained in the short-term dynamics of the slowest normal modes, which can be obtained from monomer coordinates $\{\mathbf{x}_n\}_{n=1}^N$ for every simulation frame, through a Discrete Cosine Transform (DCT),

$$\mathbf{z}_j = \sum_{n=1}^{N} \mathbf{x}_n \cos\left[\frac{\pi}{N}\left(j + \frac{1}{2}\right)n\right], \tag{15}$$

obtaining a *non-Markovian* 3D discrete time stochastic process $\{\mathbf{z}_{j,t}\}_{t>0}$, $\forall\, j \in \{1, ..., k\}$.
In such manner, there is no need to specify an *a priori* diffusive spectrum and kernel function: it would be contained in the normal modes dynamics, implicitly enclosing information about the diffusive scaling $\{\gamma_j\}_{j=1}^N$ and other important kernel features.
Normal modes role is thereupon, threefold:

1. Short-time CG MD trajectories pertaining to k modes are used to train a set of independent NAR generative models with the objective of learning the conditional distribution $p(\mathbf{z}_{j,t}|\mathbf{z}_{j,t-\Delta t}, \mathbf{z}_{j,t-2\Delta t}, ..., \mathbf{z}_{j,t-m\Delta t})$, for a fixed historical trajectory of size m, and fixed coarse-grained Δt time-step. See section 4.2.
2. We assume that normal modes act as CVs and at the same time carry sufficient information about the environment to determine the long-time behaviour. Consequently, it is reasonable that their short-time trajectories can be used to fit the resulting MSD equation derived from Eq. (13). See section 3.5.
3. Normal modes can be used to reconstruct approximate single polymer configurations with less d.o.fs, since the mapping monomer-mode is bijective (see Fig. 3). This constitutes by all means another level of *coarse-graining*.

Once the NARs are trained on short trajectories, they can auto-regressively generate the modes dynamics and if the accumulation error (*exposure bias*) is small, extrapolate the auto-correlations effectively, therefore accelerating the MD simulation. As a result, this leads to accurate estimation of statistical properties, such as diffusion coefficients and at the same time it reproduces the transient MSD curve.
Albeit the MZ formalism and the GLE is derived from a generic Hamiltonian system, we assume the same can be done with an already CG system which is inherently stochastic. This does not impact the applicability of the method, since we could follow the same procedure with an All-Atom simulation by selecting the same CG variables (monomers).

3.1. Neural auto-regressive Generative Model for non-Markovian stochastic processes

Parameter inference on *non-Markovian* stochastic processes is in general a difficult task. It is challenging therefore to reproduce and generate signals with the same statistical properties [49].

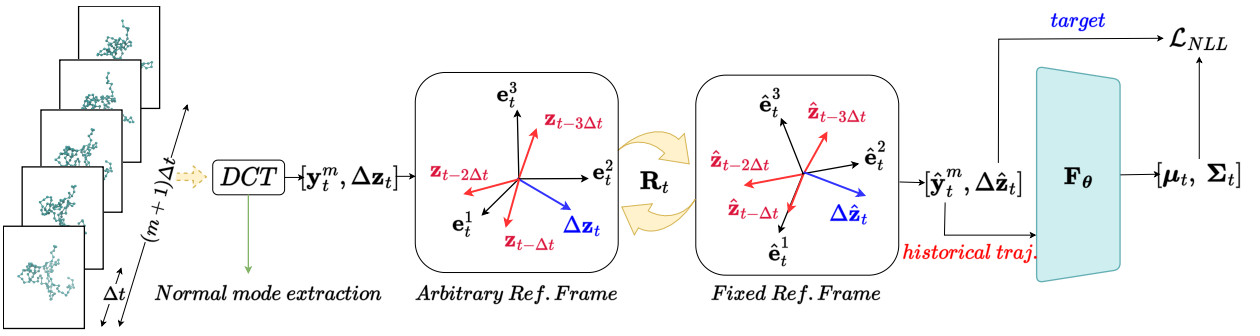

Fig. 4: NAR Training scheme. Mode j (subscript dropped) is extracted from a sequential collection of polymer configurations of length $m+1$ with a DCT. sub-trajectory of length m and target increment are rotated to a fixed reference frame, such that $\hat{\mathbf{e}}_t^1$ is always aligned with $\hat{\mathbf{z}}_{t-\Delta t}$, while $\hat{\mathbf{z}}_{t-2\Delta t}$, lies in the plane formed by $\hat{\mathbf{e}}_t^1, \hat{\mathbf{e}}_t^2$. Afterwards, the historical mode trajectory is fed to a NN, which outputs the parameters of a 3D Gaussian distribution. The target increment vector $\Delta\hat{\mathbf{z}}_t$ is used to compute the NLL.

NNs have demonstrated to be capable of modeling non-linear relationships, thanks to the universal approximation theorem [50]. Our aim is therefore to generate 3D modes dynamics $\{\mathbf{z}_{j,t}\}_{j=1}^k$ only using short MD time-series sampled each Δt and coming from an equilibrium polymer melt simulation, and capable of reproducing both long-term auto-correlations, and stationary empirical PDF.
Given historical information, we parameterized the conditional distribution with a NN, one for each mode process $\mathbf{z}_{j,t}$. Since modes are generated independently, in the following, for notational simplicity, we drop the subscript j indexing the mode number. We further discuss modes independence approximation in section 3.5.

We denote as $\mathbf{y}_t^m = \{\mathbf{z}_{t-i\Delta t}\}_{i=1}^m \in \mathbb{R}^{3\times m}$ the 3D mode historical trajectory and as $\Delta\mathbf{z}_t \in \mathbb{R}^{3\times 1}$ its successive time increment vector. Since each polymer is at equilibrium in its environment which acts a bath, it is safe to assume that the processes possess a finite size memory of order m, corresponding to a time lag $t_{max} = m\Delta t$, and that $p(\mathbf{z}_t|\mathbf{y}_t^m)$ is stationary. We propose therefore to approximate the latter with a parameterized distribution $p_\theta(\Delta\mathbf{z}_t|\mathbf{y}_t^m)$ for the time increment $\mathbf{z}_t - \mathbf{z}_{t-\Delta t}$, followed by an integration step. We observed better performances in optimizing on the discrete differential $\Delta\mathbf{z}_t$ instead of the actual process value $\mathbf{z}_t$. We argue that differencing stabilizes the learning process and is more suited in case of long memory processes, where convergence to the true stationary distribution may be very slow. The Gaussian distribution constitutes a good *ansatz*, since chains are at equilibrium and modes follow the Gibbs distribution, hence,

$$p_\theta(\Delta\mathbf{z}_t|\mathbf{y}_t^m) \sim \mathcal{N}(\Delta\mathbf{z}_t|\boldsymbol{\mu}_\theta(\mathbf{y}_t^m), \boldsymbol{\Sigma}_\theta(\mathbf{y}_t^m)) \tag{16}$$

where $\boldsymbol{\mu}_\theta : \mathbb{R}^{3\times m} \to \mathbb{R}^3, \boldsymbol{\Sigma}_\theta : \mathbb{R}^{3\times m} \to \mathbb{R}^{3\times 3}$ are two NNs parametrizing the conditional mean and the positive semi-definite and symmetric conditional covariance matrix, both depending on instances of historical trajectories. We prove the goodness of this choice in Appendix B, where we demonstrate the Gaussian nature of both the MD and generated process.
Each $\mathbf{y}_t^m$ feeds a parameterized network $\mathbf{F}_\theta$ (MLP) which encodes input-output correlations and forecasts the two distribution parameters. $\Delta\mathbf{z}_t$ is used as target of the supervised training scheme, which is illustrated in Fig. 4.

3.2. Symmetries

Symmetries can be implemented natively in different ways, from equivariant NNs [51] to graph NNs [52], the latter suited for particle simulations because symmetries can be naturally embedded in the graph formalism. [53, 54]. Our solution is much easier and is exploits the fact that we are generating an effective dynamics irrespective of the environment configuration, as we generate single polymer normal modes dynamics.
Polymer diffusion is in fact isotropic and in absence of external driving perturbations breaking spherical symmetry, the normal modes conditional distribution should be *invariant* under any arbitrary global rotation, i.e.,

$$p(\Delta\mathbf{z}_t|\mathbf{y}_t^m) = p(\mathbf{R}\Delta\mathbf{z}_t|\mathbf{R}\circ\mathbf{y}_t^m), \tag{17}$$

where $\mathbf{R} \in \mathrm{SO}(3)$, and is applied *element-wise* to $\mathbf{y}_t^m$.
This observation is crucial for generation stability because training in a fixed reference frame spares the network from learning rotations in SO(3), and avoids the need for *data augmentation*, or more complicated architectures.

The rotation matrix at each time-step, $\mathbf{R}_t = [\hat{\mathbf{e}}_t^1, \hat{\mathbf{e}}_t^2, \hat{\mathbf{e}}_t^3]$ is defined from the last three elements of $\mathbf{y}_t^m$, following the *Gram-Schmidt* orthonormalization. The result is a fixed orthonormal set spanning $\mathbb{R}^3$. In this way, optimization is performed in the same reference frame because data instances are transformed accordingly prior to training (Fig. 4). Hence,

$$\hat{\mathbf{y}}_t^m = \mathbf{R}_t \circ \mathbf{y}_t^m, \tag{18}$$

$$\Delta\hat{\mathbf{z}}_t = \mathbf{R}_t \Delta\mathbf{z}_t. \tag{19}$$

Rotations operations are invertible, allowing the auto-regressive process generation.

3.3. Loss function

Training is performed in a supervised manner, using the rotated historical trajectory as input and the consecutive mode time increment as target and optimizing over the Negative Log-Likelihood (NLL) of a 3D Gaussian distribution, which reads,

$$\mathcal{L}_{NLL} = -\log \mathcal{N}(\Delta\mathbf{z}_t | \boldsymbol{\mu}_{\theta,t}(\hat{\mathbf{y}}_t^m), \boldsymbol{\Sigma}_{\theta,t}(\hat{\mathbf{y}}_t^m))$$

Up to a constant, the NLL can be expressed as,

$$\mathcal{L}_{NLL} = \log\det\boldsymbol{\Sigma}_t + (\Delta\hat{\mathbf{z}}_t - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\Delta\hat{\mathbf{z}}_t - \boldsymbol{\mu}_t) \tag{20}$$

where $[\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t] = \mathbf{F}_\theta(\hat{\mathbf{y}}_t^m)$.
It is straightforward to see that optimization leads to training instabilities. To demonstrate this fact, one can look at the loss gradient,

$$\nabla_\theta \mathcal{L}(\theta) \propto \nabla_\theta \det\boldsymbol{\Sigma}(\theta) / \det\boldsymbol{\Sigma}(\theta).$$

This implies that the *backpropagation* signal would be sensitive to small entries in the covariance matrix, leading to exploding loss values. By replacing the LDL^T decomposition of the covariance matrix, instabilities in the loss function gradient are greatly reduced, obtaining,

$$\mathcal{L}_{NLL} = \mathrm{Tr}(\log\mathbf{D}_t) + (\Delta\hat{\mathbf{z}}_t - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\Delta\hat{\mathbf{z}}_t - \boldsymbol{\mu}_t) \tag{21}$$

where $\mathbf{D}_t$ is the diagonal matrix of the decomposition of $\boldsymbol{\Sigma}_t = \mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^T$. $\mathbf{L}_t$ is instead a uni-triangular 3×3 matrix.
Its contribution to the loss function vanishes as the log-determinant is zero, leaving only the log-determinant of a diagonal matrix, which is equivalent to the trace of the same matrix.
At the architectural level, let d be the process dimensionality. Instead of forecasting the $d(d+1)/2$ elements of the covariance matrix, $\mathbf{F}_\theta$ also outputs a single vector $[\mathbf{l}, \mathbf{d}]$ of the same size, which contains the elements constituting the lower triangular part of $\mathbf{L}_t$ and the diagonal entries of $\mathbf{D}_t$ respectively. Hence,

$$[\boldsymbol{\mu}_t, \mathbf{D}_t, \mathbf{L}_t] = \mathbf{F}_\theta(\hat{\mathbf{y}}_t^m) \tag{22}$$

Along these lines, one could employ a different distribution for approximating the process increments, provided one is able to obtain an analytic maximum likelihood objective. Furthermore, since the only other assumption is that the process is stationary, this neural auto-regressive technique could be used a broader class of bounded stochastic processes coming from various physical systems. For example, this technique could be applied to a particle undergoing an harmonic or double-well potential as long as the initial hypothesis are respected.

3.4. Auto-regressive Generation

After training, the NN block is used auto-regressively. The rotation matrix $\mathbf{R}_t$ is computed from the input historical trajectory $\mathbf{y}_t^m$, to obtain the rotated trajectory $\hat{\mathbf{y}}_t^m$. The latter is then fed to $\mathbf{F}_\theta$ and the output is used to sample $\Delta\hat{\mathbf{z}}_{j,t}$ by means of the *reparametrization trick*. The sampled time increment is then rotated back to the original reference frame of the historical trajectory $\mathbf{y}_t^m$. Given $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_3)$ one has,

$$\Delta\hat{\mathbf{z}}_t = \boldsymbol{\mu}_t + \mathbf{L}_t\mathbf{D}_t^{1/2}\boldsymbol{\epsilon}, \tag{23}$$

$$\Delta\mathbf{z}_t = \mathbf{R}_t^{-1}\Delta\hat{\mathbf{z}}_t \tag{24}$$

Following an implicit *Euler* integration scheme, a new value of $\mathbf{z}_t$ is therefore obtained,

$$\mathbf{z}_t = \mathbf{z}_{t-\Delta t} + \frac{\sigma_{\Delta z}}{\sigma_z}\Delta \mathbf{z}_t \Delta t \tag{25}$$

where $\sigma_{\Delta z}/\sigma_z$ are the empirical standard deviations and are needed for consistency, since $\Delta\mathbf{z}, \mathbf{z}$ are both standardized in pre-processing. The new input trajectory is obtained with a sliding window on the new sampled value.
Starting from an MD initial condition, $\hat{\mathbf{y}}_0^m$ and iterating the procedure, including the newly generated point in the new input sub-trajectory, the process can be extrapolated at any simulation length, reproducing useful statistical properties such as the auto-correlation function (ACF) at long times. It is important to point out that modeling the conditional distribution as a multivariate Gaussian can be restricting when approximating the conditional distribution. The latter being a skewness zero distribution, makes it difficult to reproduce the extreme regions of the PDFs.

3.5. Learning the GLE parameters

In order to model the C.o.M stochastic dynamics, we exploit the *normal modes* trajectories obtained from short MD simulations, together with the GLE solution, as written in Eq. (13). We observe that since the system is at equilibrium and diffusion is isotropic, the MSD can be fitted on 1D C.o.M and modes trajectories.
Rewriting therefore Eq. (13) as a 1D finite difference equation, squaring and taking expectations with respect to stationary equilibrium distribution, we derive the following regression equation,

$$\mathbb{E}[\Delta X_{k,t}^2] = \alpha^2 t + \beta^2 \sum_{j=1}^{k} \mathbb{E}[\Delta z_{j,t}^2] \tag{26}$$

where we imposed modes statistical independence, $\mathbb{E}[z_{i,t}z_{j,t}] = \sigma_{ij}^2\delta_{ij}$. This is a good approximation as shown by the empirical correlation matrix, displayed in Fig. 5.
We can recognize on the left hand side the C.o.M MSD, while on the right hand side the modes MSD, which are all bounded being *mean reverting* processes; the Brownian motion MSD is instead well known for being linear in time.
Replacing the expectation operator $\mathbb{E}[...]$ with the empirical average $\langle ... \rangle$ with respect to the equilibrium distribution, we can find the best parameters in the least square sense,

$$\alpha_k^*, \beta_k^* = \arg\min_{\alpha,\beta} \left\{ \sum_{t\in\mathcal{P}} \left(\langle \Delta X_{k,t}^2 \rangle - \alpha^2 t - \beta^2 \sum_{j=1}^{k} \langle \Delta z_{j,t}^2 \rangle \right)^2 \right\}, \ s.t.\ \alpha, \beta \geq 0 \tag{27}$$

where $\mathcal{P}$ is the set of data-points indexes used for regression.
The parameter inference problem is reduced to a simple multivariate regression between the modes and C.o.M MSD, with an arbitrary number of regression points to be decided. One can observe the best solution depends on how many normal modes k are used. More details on the latter can be found in section 4.5.
By combining the solution of Eq. (27) with the NAR generative model for normal modes constitutes the ML-GLE, a stochastic generator for the effective polymer dynamics whose generation protocol is summarized in Alg.1.

4. Numerical experiments

4.1. Homo-polymer melt CG simulations

We hereby describe how the reference polymer melt MD simulation was performed. Polymer melts are viscoelastic materials notoriously difficult to simulate because their enchained structure makes the dynamics very slow. In this case, initial equilibration can be quite expensive as well, increasing further with system size.
In order to test the ML-GLE framework robustness in reproducing single polymer dynamics, we performed CG MD simulations of a 100 *cis-Polybutadyene* identical polymers of $N = 100$ monomers each, in a wide range of temperatures. Polybutadyene, also called *Butadiene Rubber* is used ubiquitously in tire manufacturing, but also in combination with other plastics, as it improves their mechanical properties. Moreover, it has recently been used as propellant in solid rocket boosters.

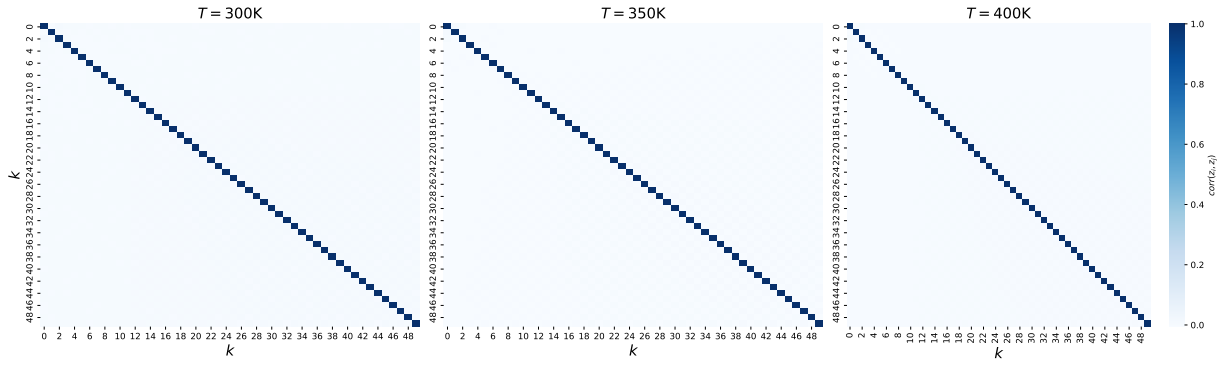

Fig. 5: Modes correlation matrices at $T = 300\,\mathrm{K}$ (Left), $T = 340\,\mathrm{K}$ (Center), $T = 400\,\mathrm{K}$ (Right), for the slower 50 modes. The correlations are computed from MD configurations and are averaged over all polymer configurations. Cross mode correlations are negligible and this justifies the independence approximation, which allows to derive the regression Eq. (26). Correlations may still be present especially for faster modes, which are related to the local polymer melt atomistic structure.

Algorithm 1 ML-GLE: Generation

Require: $\mathbf{y}^m_{j,t_{max}+\Delta t}\ \forall j \in \{1,..,k\},\ \mathbf{X}_0,\ N_{gen},\ \alpha^*_k,\ \beta^*_k,$ ▷ Initial MD trajectory, Initial C.o.M position, fit parameters

for $t \in \{t_{max}+\Delta t, ..., T_{gen}\}$ **do**

for $j \in \{1, ..., k\}$ **do**

$\hat{\mathbf{y}}_{j,t} \leftarrow \mathbf{R}_{j,t} \circ \mathbf{y}_{j,t}$ ▷ Sub-trajectory rotation

$[\boldsymbol{\mu}_{j,t}, \mathbf{D}_{j,t}, \mathbf{L}_{j,t}] \leftarrow \mathbf{F}_{j,\theta}(\hat{\mathbf{y}}^m_{j,t})$ ▷ Input encoding with MLP

$\boldsymbol{\Sigma}_{j,t} \leftarrow \mathbf{L}_{j,t}\mathbf{D}_{j,t}\mathbf{L}^T_{j,t}$ ▷ Covariance matrix reconstruction

$\Delta\hat{\mathbf{z}}_{j,t} \sim \mathcal{N}(\boldsymbol{\mu}_{j,t}, \boldsymbol{\Sigma}_{j,t})$ ▷ Sampling in the fix ref. frame

$\Delta\mathbf{z}_{j,t} \leftarrow \mathbf{R}^{-1}_{j,t}\Delta\hat{\mathbf{z}}_{j,t}$ ▷ Rotation back to the input ref. frame

$\mathbf{z}_{j,t} \leftarrow \mathbf{z}_{j,t-\Delta t} + (\sigma_{\Delta\mathbf{z}_j}/\sigma_{z_j})\Delta\mathbf{z}_{j,t}\Delta t$ ▷ Implicit Euler integration

$\mathbf{y}_{j,t+\Delta t} \leftarrow (\mathbf{y}_{j,t} \setminus \{\mathbf{z}_{j,t-m\Delta t}\}) \cup \{\mathbf{z}_{j,t}\}$ ▷ Sub-trajectory update

end for

$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Delta t)$

$\mathbf{B}_t \leftarrow \mathbf{B}_{t-\Delta t} + \boldsymbol{\xi}$ ▷ 3D Brownian motion integration

$\mathbf{X}_t \leftarrow \mathbf{X}_{t-\Delta t} + \alpha^*_k\mathbf{B}_t + \beta^*_k \sum_{j=1}^k \sigma_{z_j}\mathbf{z}_{j,t}$ ▷ 3D C.o.M dynamics integration

end for

The simulations are done using Dissipative Particle Dynamics (DPD). Upon selection of monomer CG variables, DPD equations can be formally expressed as follows $\forall n \in \{1, ..., N\}$,

$$\begin{cases} \dot{\mathbf{x}}_n = \mathbf{v}_n = \mathbf{p}_n/m \\ \dot{\mathbf{p}}_n = \mathbf{F}^C_n + \mathbf{F}^D_n + \mathbf{R}_n \end{cases} \tag{28}$$

where $\mathbf{F}^C_n, \mathbf{F}^D_n, \mathbf{R}_n$ are respectively the conservative (bonded and non-bonded) forces acting on monomer n, the dissipative forces and a Gaussian uncorrelated noise. Fixing a radial distance cutoff r_d, the dissipative and random force can be expressed as pairwise contributions from neighbouring CG particles.
Denoting as $\mathcal{S}_n(r_d)$ and $\mathcal{B}_n$, respectively, the set of non-bonded neighbours at distance r_d and the set of bounded neighbours for monomer n, one has

$$\mathbf{F}^C_n = \sum_{m\in\mathcal{B}_n} \mathbf{f}^C_{m\to n}(\mathbf{r}_{mn}) + \sum_{m\in\mathcal{S}_n} \mathbf{f}^C_{m\to n}(\mathbf{r}_{mn}), \quad \mathbf{F}^D_n = \sum_{m\in\mathcal{S}_n} \mathbf{f}^D_{m\to n}(\mathbf{r}_{mn}), \quad \mathbf{R}_n = \sum_{m\in\mathcal{S}_n} \mathbf{f}^R_{m\to n}(\mathbf{r}_{mn}) \tag{29}$$

Intra-molecular conservative forces, for bonds and angles, are derived from the following harmonic potentials,

$$E^b_{mn} = k_{mn}(r - r^0_{mn})^2, \quad E^a_{mnl} = k_{mnl}(\theta - \theta^0_{mnl})^2, \tag{30}$$

where r^0_{mn} is the equilibrium distance of any monomer pair mn and θ^0_{mnl} is the equilibrium angle between any three consecutive monomers mnl. k_{mn} and k_{mnl} are the respective elastic spring constants. The non-bonded conservative interactions are instead tabulated pair potentials obtained from cubic spline interpolation using 10 sampling distances.

The non-conservative dissipative force is the classical DPD additive force,

$$\mathbf{f}^D_{m\to n} = -\gamma\left(1-\frac{r}{r_d}\right)^2(\mathbf{v}_{mn}\cdot\mathbf{e}_{mn})\mathbf{e}_{mn} \tag{31}$$

where γ is the friction coefficient, $\mathbf{v}_{mn} = \mathbf{v}_m - \mathbf{v}_n$, is the relative velocity and $\mathbf{e}_{mn}$ is the unit vector directing monomer m and n. The parameters of the dissipative forces and bounded conservative forces are estimated with the statistical trajectory matching (STM) [55], a Bayesian technique which exploits *downsampled* trajectories of an *All-Atom* (AA) simulation of the same system at $T = 300\,\mathrm{K}$.

The STM technique allows therefore to estimate the equilibrium values r^0_{mn}, θ^0_{mn}, the constants k_{mn}, k_{mnl} and the friction coefficient γ, assumed identical for all monomers. Finally, the additive random force takes the form,

$$\mathbf{f}^R_{m\to n} = \sigma\left(1-\frac{r}{r_d}\right)\frac{u}{\sqrt{\delta t}}\mathbf{e}_{mn} \tag{32}$$

The fluctuation-dissipation theorem imposes $\sigma = \sqrt{2k_BT\gamma}$, while u is a standard Gaussian random number. The forces are integrated using the velocity Verlet algorithm with a simulation time-step of $\delta t = 50\,\mathrm{fs}$ and periodic boundary conditions (PBCs) for a cubic simulation box of length $l = 10\,\text{Å}$. We performed 11 simulations using the software LAMMPS and run on a CPU cluster, at different temperatures from 300 K to 400 K (snapshot in Fig. 1), at intervals of 10 K for $T_{sim} = 2\times10^8\delta t$.

These number of steps are sufficient to observe the normal diffusive regime and estimate the diffusion coefficient on all simulated temperatures. However, even reducing the overall d.o.fs with the CG scheme, these simulations are very expensive and ML-GLE speeds-up substantially single polymer effective dynamics by several orders of magnitude.

4.2. Dataset preparation

Simulated trajectories are *downsampled* on the fly with a time-step of $\Delta t = 2\times10^3\delta t = 100\,\mathrm{ps}$. This is justified by the fact that we are trying to learn a coarse-grained C.o.M integrator for the modes dynamics. Δt is chosen such that its resulting C.o.M MSD at short times is subdiffusive, meaning the ballistic regime is resolved and thereby ensuring the validity of the zero-mass limit GLE description.. Full simulation length can be thus expressed in units of Δt, hence $T_{sim} = 10^5\Delta t$.

For all simulated temperatures, normal modes are obtained from CG monomer coordinates $\{\mathbf{x}_i\}_{i=1}^N$ for all polymers, with a DCT, as described in section 3. Performing this for each sampled snapshot, we obtain a discrete 3D non-Markovian stochastic process, one for each mode. Training is performed with only 1 % of the full trajectory length, hence $T_{train} = 10^3\Delta t$. For validation, we used subsequent trajectories of length $T_{val} = 6\times10^2\Delta t$, while the rest of the dataset is used to compare the long-time behaviour of the generated trajectories (MSD, ACFs).

Sub-trajectory dataset is obtained as follows:

1. Modes data is properly standardized, $(\mathbf{z}_j - \langle\mathbf{z}_j\rangle)/\sigma_{z_j}$, where $\langle\mathbf{z}_j\rangle = \mathbf{0}$, $\forall\, j\in\{1,...,N-1\}$, since every polymer is supposed to be at equilibrium with its environment. σ_{z_j} are pre-computed from data up to T_{train}. The dataset is then splitted in train set and validation set.
2. Only the slowest $k = 12$ modes trajectories are retained from monomer coordinates. The rationale behind this choice is elucidated in section 4.5.
3. Sub-trajectory training and validation datasets are obtained with a lag-1 sliding window of length $m+1$ up to T_{train} and T_{val} respectively and for each of the 100 polymer trajectories. We obtain the target time increment by differentiation over the last sub-trajectory element.

Ultimately, we exploit the fact that the physical system is made of homo-polymers, meaning their marginal conditional distributions are equal and obtaining consequently a larger pool of different sub-trajectories that are used for *mini-batch* training, after random shuffling. Dataset pre-processing, post-processing, training and generation codes were all implemented using **PyTorch** ML library and executed on a GPU server with two **NVIDIA** A100 accelerators. Training data up to T_{train} for all temperatures is available at https://huggingface.co/datasets/gian-michele/meltBR.

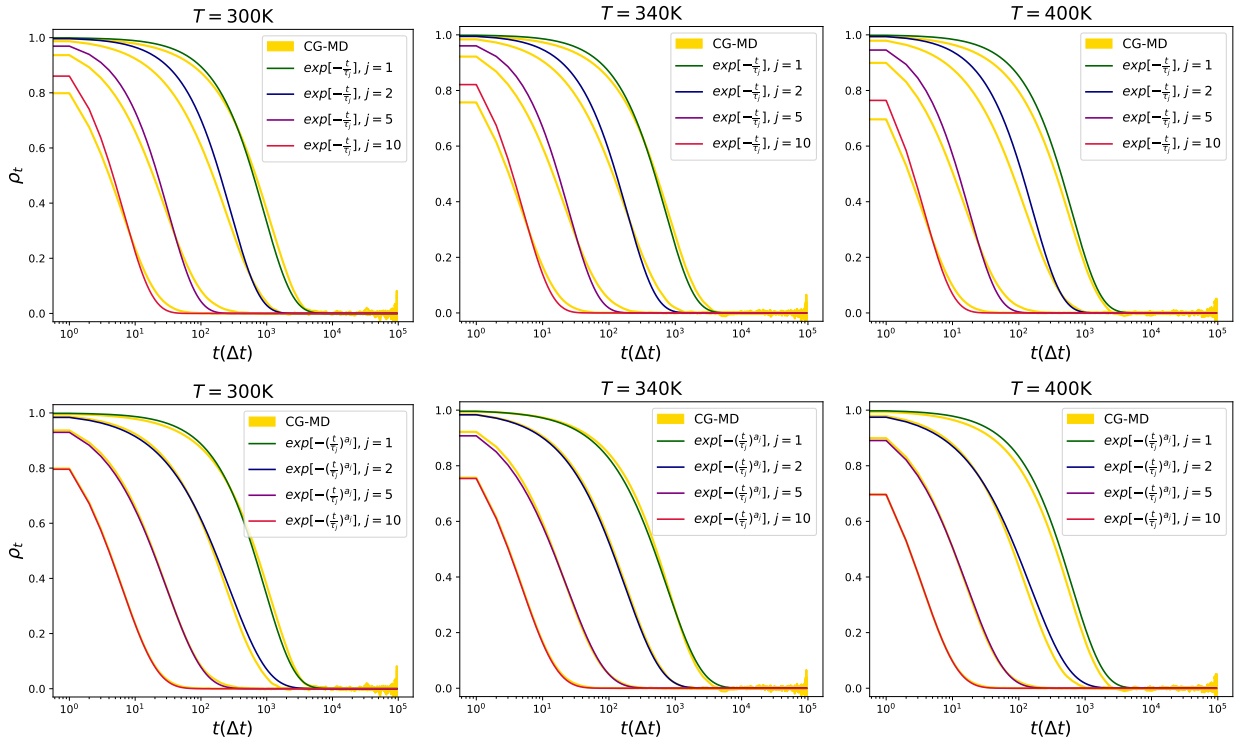

Fig. 6: Exponential (**Top**) and Stretched exponential (**Bottom**) curve fit of MD normal modes NACFs for $T = 300$ K (**Left**), $T = 340$ K (**Center**) and $T = 400$ K (**Right**). The number of NACF data-points used for both fits are 10^3 for each mode (*1st*, *2nd*, *5th*, *10th*), corresponding to T_{train}.

4.3. Normal modes dynamics beyond exponential relaxation

In this section we demonstrate the inadequacy of the *Markovian* approximation when dealing with single polymer relaxation behaviour, thus justifying the need for data-driven techniques capable of reproducing memory effects, without performing full-size MD simulations.

As stated in section 3, the GLE with integrable kernel is capable of reproducing any TAD curve with a power-law diffusive spectrum, provided the long-time relaxation behaviour is known, by means of a set of OU processes. This constitutes a real problem when dealing with realistic polymer materials simulations, because single polymer dynamics is evidently *non-Markovian* and in addition the power-law kernel can be a restrictive choice with respect to the space of integrable functions.

In the following, we demonstrate the deviations of MD single polymer normal modes from exponential relaxation characterizing OU processes, which are present in the GLE solution, as shown in Eq. (10). In Fig. 6, we fit the MD normal modes Normalized ACFs with an exponential function with the same amount of data used for training (i.e. trajectory length, N_{train}). The latter is not able to reproduce the correct relaxation suggesting that OU processes are not suitable to describe normal modes dynamics.

Relaxation in disordered media is often modeled with the *Kohlrausch* function, also known as the stretched exponential. This type of relaxation can describe different phenomena typical of complex systems, from polymer dielectric spectra [56] to MRI signals in the brain [57], dynamical heterogeneities [58], long-term correlations in spin-glass systems [59] and quantum diffusion [60]. Given $a \in (0, 1)$ and $\tau > 0$, the *Kohlrausch* function can be defined as follows,

$$f(t) = C_a e^{-(t/\tau)^a} \tag{33}$$

where the pre-factor C_a depends in general on a. The case $a = 2$ yields the Normal distribution. In Fig. 6 we show the ACF fit at different temperatures, exploiting the stretched exponential and obtaining an important improvement, demonstrating the non trivial behaviour of modes dynamics. Even in this case, noticeable deviations are present both at short times and in their tails behaviour, especially in the slower modes, suggesting that a more complex description is needed. Furthermore, even assuming the *Kohlrausch* function as a good relaxation model, generating

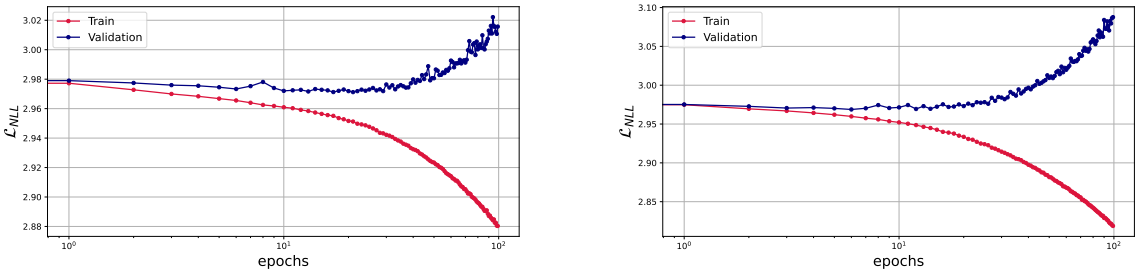

Fig. 7: NLL training and validation curves at $T = 300\,\mathrm{K}$ for $1st$ (Left) and $2nd$ (Right) normal modes. Training was performed with a 128 batch-size for 100 epochs, while validation with 256. The best epoch is chosen as the validation loss minimum.

auto-regressively from the corresponding *non-Markovian* stochastic differential equation (SDE) is a difficult task. Consequently, employing NNs seems a viable choice in order to estimate the *coarse-grained* integrator and to access the modes relaxation times, approximating the ACFs and also allow their long-term extrapolation.
Furthermore, generating normal modes trajectories has several advantages: it allows for simulate an high fidelity single polymer effective dynamics and has the potential to be used in combination with more advanced generative methods in order to construct better CG schemes.

4.4. NAR training results for Normal modes

In order to harness the capabilities of the GLE to describe and reproduce the C.o.M TAD dynamics, we propose to endow it with a NAR generative model for the modes dynamics, trained on short MD trajectories.

3D historical trajectories of length $m = 128$ are encoded, after flattening, with a Multi-Layer Perceptron (MLP) made of 2 hidden layers with a total of 512 neurons, to a latent vector of size 12. This provides the input for two separate dense layers outputting the conditional mean vector and elements of the LDL^T decomposition. Training is executed with Adam optimizer with learning rate $\eta = 10^{-4}$. The target fluctuation $\Delta\mathbf{z}_t$ is then used to evaluate the NLL.
The free parameter t_{max} can be empirically justified by looking at the $1st$ mode empirical ACF: we obtained best results for a memory size corresponding to roughly 20 % ACF decrease. An additional heuristic on this is provided in section 4.5. For each mode, the best epoch is chosen such that it minimizes the validation loss: in Fig. 7 one can clearly see that over-fitting is already apparent after a few dozens epochs for the first mode and appearing earlier for faster modes. It seems that faster decaying correlations are easier to learn, reaching therefore validation minima sooner. More training loss analysis can be found in Appendix C.
Furthermore, exploiting modes independence and given the small NNs size, for each mode, training and generation can potentially be performed in parallel. *Parallelization* over modes depends on specific use cases and hardware availability, other than the number of modes one wishes to include in the model.
The auto-regressive generation results can be appreciated by comparing the normalized auto-correlations (NACFs) between the synthetic generated trajectories and the MD ones. For each temperature, modes stochastic dynamics are generated with Alg. 1 up to $T_{gen} = T_{sim}$, for approximately 10^5 steps, from an ensemble of 300 MD initial historical trajectories. In Fig. 8, we show some preliminary examples trajectories coming from MD and NAR model for different normal modes. Notice how higher modes are characterized by smaller fluctuations.

Given the length of the generated signals, brute force empirical auto-correlation computation is impractical, having a $O(n^2)$ complexity. Exploiting the *Khinchin–Kolmogorov* theorem for stationary stochastic processes, one can compute the same auto-correlation employing two Fast-Fourier Transforms (FFTs), and a computational cost of $O(n \log n)$, where n is the trajectory length. Confidence intervals (confidence intervals) for ACFs are instead computed following Mélard and Roy [61], where the standard error (SE) at lag k relative to an underlying *moving average* stochastic process is expressed as,

$$\sigma^2_{SE,k} = \frac{1}{n}\left(1 + 2\sum_{i=1}^{k-1}\rho_i^2\right) \tag{34}$$

where $\rho_i = \langle z_0 z_i\rangle/\sigma_z$ is the lag i normalized correlation estimator for the mode z.
Example NAR generated trajectories at $T = 300\,\mathrm{K}$ are shown in Fig. 8, while Normalized ACFs for some modes

at different temperatures are shown in Fig. 9, together with the empirical PDFs for the Δz_j, z_j. The dynamical model shows remarkable capabilities in extrapolating far beyond what has been trained on. The NACFs are correctly reproduced, with a net gain in computational costs with respect to full-length MD simulations. The main statistical features are therefore reproduced only training on sub-trajectories of length t_{max}, and despite not having enforced any of those features directly in the loss function. Furthermore, the modes auto-correlation deviate from classical exponential relaxation and the NNs are able to reproduce well the NACFs, outperforming the *Kohlrausch* function ACF fit at short times.

Although with low probability, the slower normal modes PDFs tails exhibit small out-of-distribution sampling that deviate from normality. This behaviour is less relevant at higher temperatures where memory decays faster. Some examples are provided in Appendix D.

A probable diagnosis for this problem resides in the spatial symmetry of the Gaussian approximation when sampling on extreme values. Limited observational data on low probability regions could also affect the result.

Architectural limitations of the MLP in handling sequences could also contribute, as well as to some other minor flaws: one can notice that if the historical trajectory's length is close to the mode relaxation time, finite size effects produce artifacts in the NACFs.

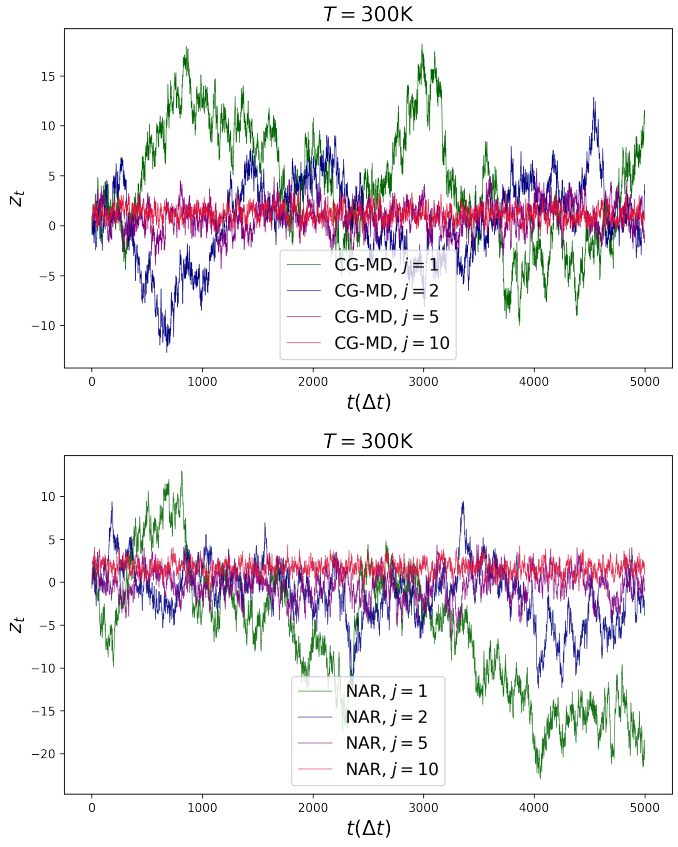

Fig. 8: (**Top**) Example modes trajectories from the MD simulation at $T = 300\,\mathrm{K}$. 1*st* mode (*green*), 2*nd* (*navy*), 5*th* (*purple*) and 10*th* (*red*). (**Bottom**) Example modes trajectories generated with the NAR model trained on trajectories until $T_{fit} = 10^2 \Delta t$ at $T = 300\,\mathrm{K}$. One can observe how slower modes exhibit larger fluctuations with longer time before reverting to their mean, which is identically zero. Faster modes fluctuate instead rapidly around zero. In both examples, the first value was subtracted for all modes, so that they would be aligned to $z_0 = 0$.

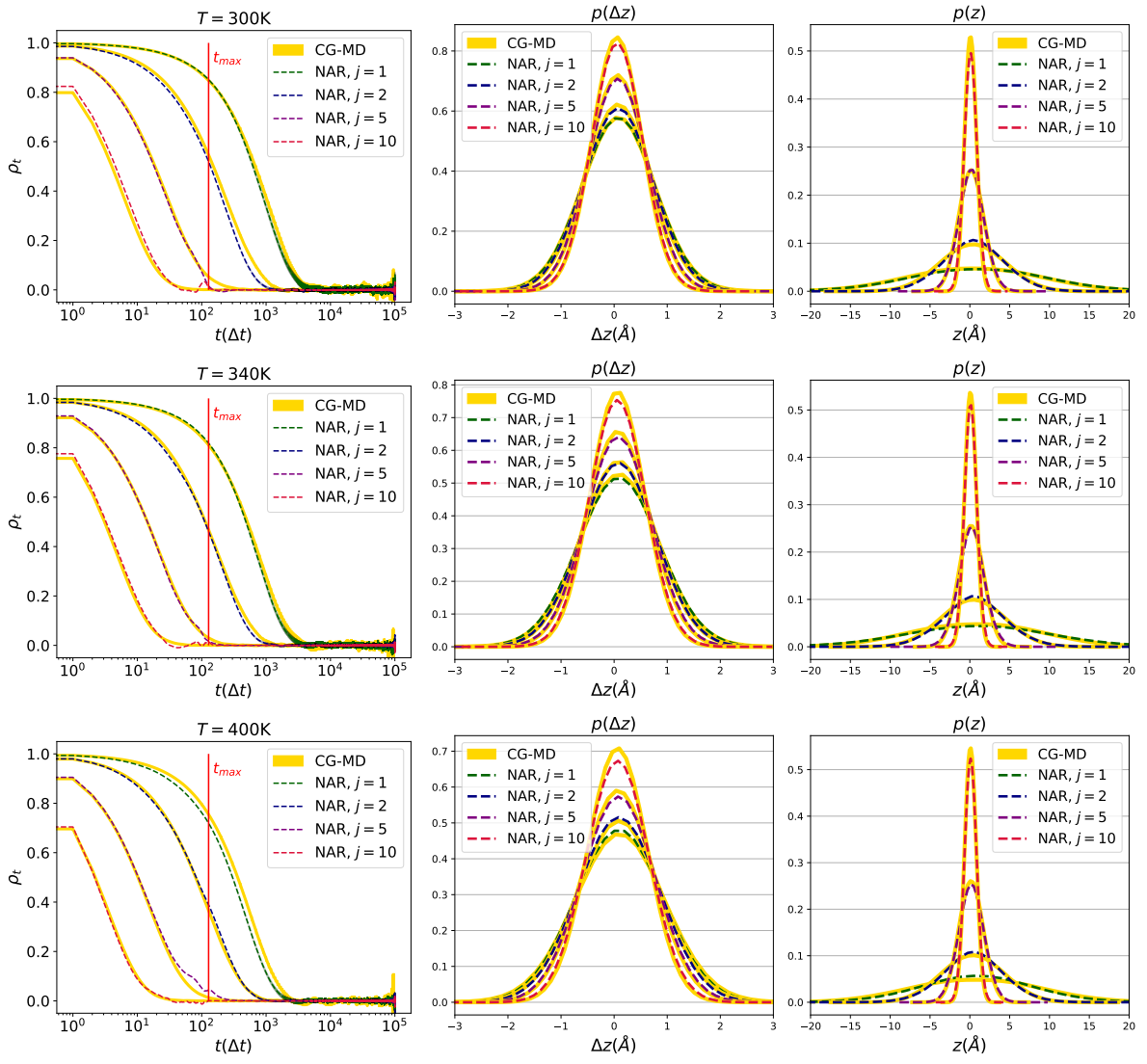

Fig. 9: (**Left**) NACFs for $T = 300\,\mathrm{K}$ (**top**), $T = 340\,\mathrm{K}$ (**middle**) and $T = 400\,\mathrm{K}$ (**bottom**) estimated from generated trajectories of length 10^5 steps with the trained NAR model trained on short trajectories of length $t_{max} = 128\Delta t$, while the full amount of training data corresponds to a trajectory length of $T_{train} = 10^3\Delta t$, 1 % of the simulated steps. The 99 % confidence intervals are computed using Eq. (34). Notice how for faster modes, the confidence intervals become progressively narrower and comparable to the reference MD ones, converging to mean value. (**Center**) The target Δz PDFs are well reproduced as one can see from the empirical PDFs for the same temperatures. (**Right**) Empirical z PDFs are reproduced as well, even if the information about the stationary distribution and ACFs was not directly embedded in the loss function.

4.5. GLE validation in reproducing TAD

To validate the GLE solution in reproducing the MSD curves relative to TAD, we first use the k slowest modes and the C.o.M MD trajectories to fit the regression Eq. (27).

The data-points used for regression are obtained computing the empirical averages at different lag times $t \in \mathcal{P}$, $\langle \Delta X^2_{k,t} \rangle$, $\langle \Delta z^2_{j,t} \rangle$, where $X_{k,t}$ is the GLE solution with k modes processes.

In order to maintain a gain in computational costs, the parameter estimation should be successful when computing the empirical averages with the same or less amount of data used to train the NAR, corresponding therefore to the trajectory length, $T_{train} = 10^3\Delta t$. At the same time, the latter should be such that the transient behaviour is not yet observed. We denote thus as T_{fit} as the trajectory length over which $\langle \Delta X^2_{k,t} \rangle$, $\langle \Delta z^2_{j,t} \rangle$, $\forall\, j \in \{1, ..., k\}$ are computed.

Regarding the set of regression data-points we use only three, $\mathcal{P} = \{1, t_{max}/2, t_{max}\}$, where t_{max} is the same sub-trajectory length chosen as historical input in the NAR model. From the GLE regression point of view, the choice

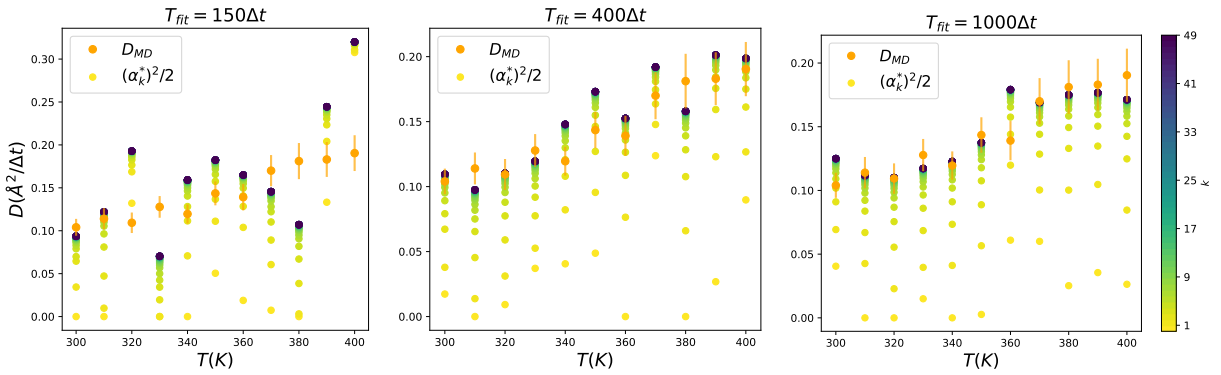

Fig. 10: Comparison between the diffusion coefficient as estimated from the reference MD trajectories, D_{MD} and the ones obtained fitting the GLE solution at an increasing number of modes, $D_{GLE} = (\alpha_k^*)^2/2$, shown by the colorbar on the right side, for all simulated temperatures. Results are shown for 3 different increasing values of T_{fit}, used to compute $\langle \Delta X_{k,t}^2 \rangle$, $\langle \Delta z_{j,t}^2 \rangle$. Error bars on D_{MD} show the 95% confidence interval computed over the 100 available polymer chains, while D_{MD} is obtained following Eq. (35) with $t_1 = 3 \cdot 10^4 \Delta t$, $t_2 = 6 \cdot 10^4 \Delta t$ for a total of $3 \cdot 10^4$ points.

of t_{max} can also be justified in an heuristic manner: we argue in fact that a small value would not capture properly the overall dynamical evolution of the slowest modes, while a value close to the transition would bias the parameter estimation.

Regression results are shown in Fig. 10, where we compare the C.o.M. diffusion coefficient obtained from the MD simulation, D_{MD}, and the one obtained from the best fit α_k^* for an increased number of modes, for different values of T_{fit}.

For the former, estimation is done choosing a time interval $[t_1, t_2]$ completely included in the diffusive regime. In particular t_1 is chosen sufficiently far from the transition so that estimation is not biased and t_2 not excessively close to the last points available in time so not be affected by sampling errors. The estimated diffusion coefficient is computed thus as follows,

$$D_{MD} = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \frac{\langle \|\mathbf{X}_{k,t} - \mathbf{X}_{k,0}\|_2^2 \rangle}{6t} \tag{35}$$

The latter is straightforward from Eq. (26). Hence, $D_{GLE} = (\alpha_k^*)^2/2$.

Fig. 10 proves that only the slower modes are needed for the GLE equation to predict the diffusion coefficient. D_{GLE} converges in fact rapidly with k and settles close to D_{MD} but with some consistent errors.

The source of these errors can be explained with the slow exploration of phase space: although polymer melt simulations are *ergodic* in the long term, they are considerably sensible to initial conditions as well. Multiple independent simulations or longer trajectories would entail a smaller error on the empirical averages about the true expectations as wider phase space regions are explored.

In order to substantiate this claim, we set hereinafter $k = 12$, and evaluate the prediction error varying the trajectory length T_{fit}. In Fig. 11, the MSE between D_{MD} and D_{GLE} for different values of T_{fit} is shown, averaged over all available temperatures. The MSE decreases quickly as T_{fit} increases, remaining substantially unaltered afterwards, but the error dispersion over temperatures continues to decrease, increasing therefore the prediction consistency over temperatures.

This is a significant result because it confirms the validity of the GLE solution in predicting the diffusion coefficient and furthermore proves that only a few single polymer CVs are relevant for the long-term behaviour.

Finally, the MSD curves can be obtained by feeding the modes trajectories to Eq. (13), and we can evaluate if this model yields consistent results and reproduces TAD.

The brute-force approach for the MSD computation is expensive for long trajectories. Exploiting the same theorem used for ACFs calculation, one can compute the MSD more efficiently by using FFTs. To show this, given a trajectory length T, we can express the squared displacement at lag m, ΔX_m^2 as,

$$\Delta X_m^2 = \frac{1}{T-m} \sum_{i=0}^{T-m-1} (X_{i+m} - X_k)^2 = \frac{1}{T-m} \sum_{i=0}^{T-m-1} (X_{i+m}^2 + X_k^2) - \frac{2}{T-m} \sum_{i=0}^{T-m-1} X_{i+m} X_k \tag{36}$$

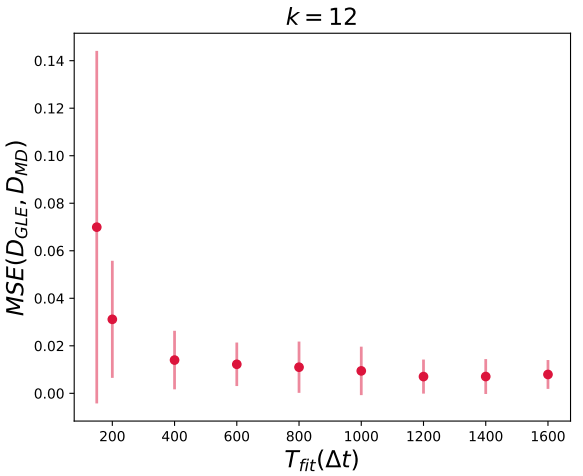

Fig. 11: Mean Square Error (MSE) between D_{MD} and D_{GLE}, for $k = 12$, over 11 temperatures. The 95% confidence intervals were estimated using the t-student quantiles. This avoids underestimation of error bars when the number of samples is low. The plot demonstrates the increase in prediction accuracy at increasing T_{fit} with a simultaneous decrease in the uncertainty on the MSE.

The second term is the position auto-correlation and can be efficiently calculated (see section 4.4). The first term can be instead computed recursively as shown in [62]. In Fig. 12, we can see that the fit GLE solution reproduces well the MSD and the positional distribution. Overall, normal modes short term correlations can thus be simultaneously exploited to integrate the C.o.M dynamical equation, gaining knowledge about the asymptotic diffusive behaviour and simulate a CG single polymer effective dynamics, through the NAR generative model.
Although normal modes work well for the observed temperature regime, approaching the glass transition these CVs may fail in describing the correct C.o.M scaling, as they are obtained with a simple linear mapping from monomer coordinates. A meaningful GLE description could be obtained by employing non-linear CVs and relaxing the independent modes assumption.

4.6. ML-GLE: Generation

Once the best NAR model is chosen for each mode, it can be used auto-regressively to generate modes dynamics and together with the fit GLE solution, generate completely synthetic C.o.M data, extrapolating its long time behaviour which is inaccessible form short time MD trajectories.
3D modes dynamics is generated starting from an initial MD historical trajectory of length m. After that, the predicted parameters are used to sample a new time increment vector $\Delta\hat{\mathbf{z}}_{j,t}$ from a Gaussian distribution which is rotated back to the original reference frame and integrated following the implicit Euler scheme as in Alg. 1.
One can appreciate the statistical relevance of our results in Fig. 13, where we show the MSD computed from generated trajectories at different temperatures, observing a very good agreement between MD and ML-GLE MSD curves. ML-GLE is able to extrapolate the dynamics up to 10 µs, sufficient to observe TAD, while only overall trajectories used for training are of length 0.1 µs, corresponding to T_{fit}.
In Fig. 14 we compare the estimated diffusion coefficients from MD simulations, the ones obtained from ML-GLE, and the fitted parameter from short trajectories that we used to generate the C.o.M dynamics. One can clearly see how the finite size trajectories affect their estimation. The ML-GLE diffusion coefficient should converge to D_{GLE} over long times but the degree of convergence is not the same for all temperatures.
As already discussed in section 4.5, sensitivity to initial conditions can impact the GLE predictive capabilities, and longer trajectories or independent simulations could allow to improve the prediction error, although that would imply a smaller gain in computational costs. Several other sources of errors could affect the final result. GLE fitting and NAR training are both performed assuming independence and identical distributed processes.
De facto, polymer interactions in the simulation box make the independence assumption invalid: single polymers are not necessarily independent, as they come from the same system. Small correlations may be present when training on sub-trajectory *batches*, effectively reducing dataset diversity.
This issue tends to be less relevant for larger systems, as more statistics on distant less correlated single polymers

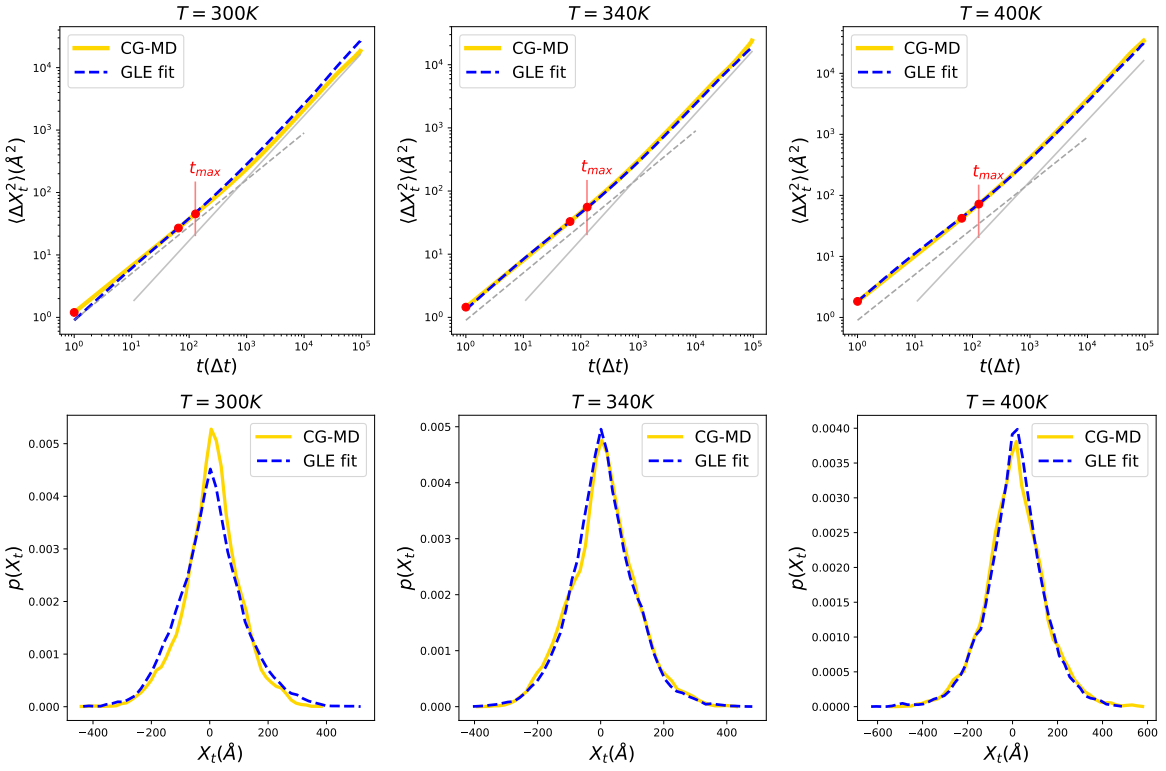

Fig. 12: (**Top**) MSD curve comparison between the MD and the fit GLE solution for $T = 300\,\mathrm{K}$ (**Left**), $T = 340\,\mathrm{K}$ (**Center**) and $T = 400\,\mathrm{K}$ (**Right**). The GLE parameters are obtained from short trajectories of length $T_{fit} = 10^3 \Delta t$. Independent Brownian processes and mode trajectories until $T_{sim} = 10^5 \Delta t$ are fed to the fit GLE solution in order to generate the C.o.M dynamics. (**Bottom**) Positional PDFs at $T_{sim} = 10^5 \Delta t$ are obtained from the C.o.M diffusive dynamics with $\mathbf{X}_0 = 0$. This PDF is of course dependent on the trajectory length, as it will asymptotically explore the whole position space.

are available. Consequently, ML-GLE is also scalable, as increasing system size would imply more trajectory data, consequently requiring even shorter simulations.
For completeness, in Table 1 we show the numeric results of the diffusion coefficients estimation for all available temperatures.

	T(K)										
$D(\AA^2/\Delta t)$	300 K	310 K	320 K	330 K	340 K	350 K	360 K	370 K	380 K	390 K	400 K
MD	0.104	0.114	0.109	0.127	0.120	0.143	0.140	0.170	0.181	0.183	0.194
Fit GLE	0.122	0.109	0.107	0.114	0.120	0.135	0.177	0.167	0.173	0.175	0.170
ML-GLE	0.141	0.116	0.116	0.113	0.113	0.140	0.163	0.164	0.155	0.194	0.161

Table 1: Diffusion Coefficients as estimated from full length MD simulation in comparison with the parameter estimated from the GLE model and the ML-GLE generated modes dynamics with Neural Networks. The parameter fit and subsequent generation was performed with $k = 12$ normal modes.

5. Conclusion

In this work we presented ML-GLE, a machine learning framework which demonstrates the predictive capabilities of the Generalized Langevin Equation when modeling complex dynamics happening at different timescales, such as *transient anomalous diffusion*. This was possible thanks to a relationship existing in the GLE framework, between slow-varying d.o.fs and long-term dynamics, determined by the tail of the kernel function.
In the case of polymer melts, although the slow variables are assumed to be uncorrelated, they are non-Markovian

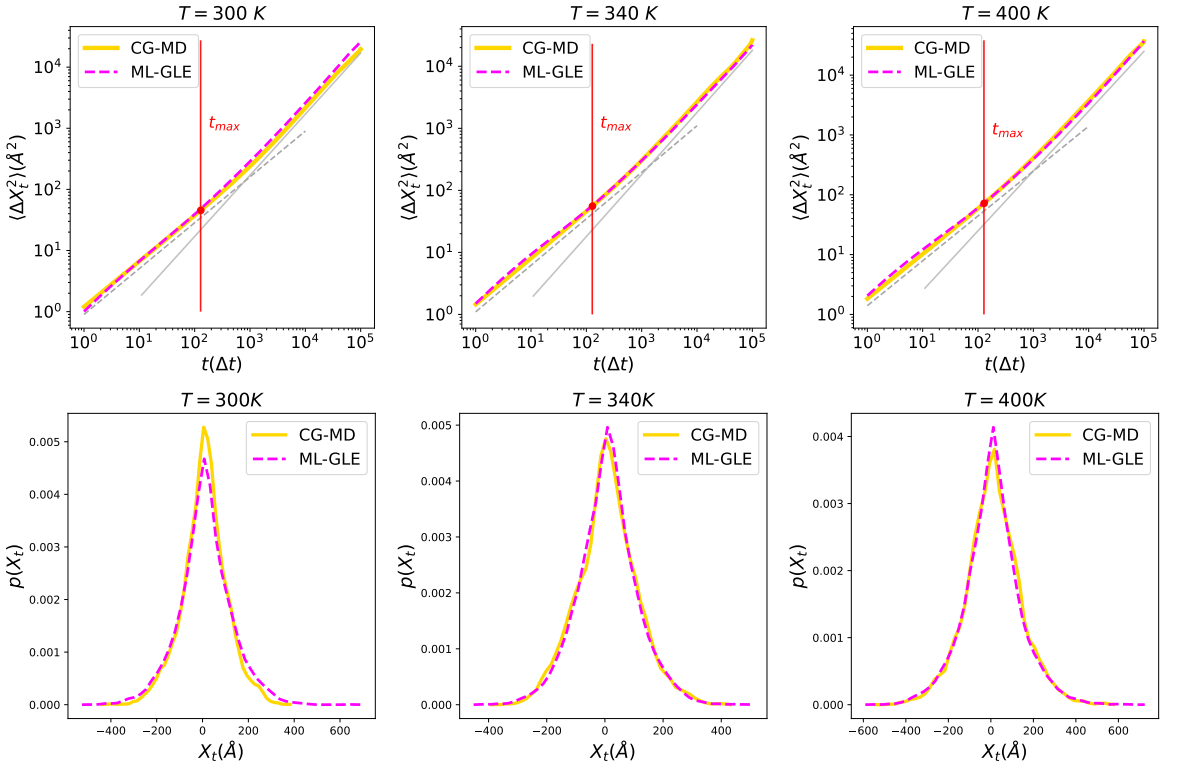

Fig. 13: (**Top**) MSD curve comparison between the MD and the C.o.M dynamics generated with ML-GLE for $T = 300\,\mathrm{K}$ (**Left**), $T = 340\,\mathrm{K}$ (**Center**) and $T = 400\,\mathrm{K}$ (**Right**). Starting from 300 initial sub-trajectories of length $t_{max} = 128\Delta t$, the modes dynamics is generated with the NAR which is trained on data equivalent $0.1\,\mu\mathrm{s}$ and generated trajectories are extrapolated therefore until $10\,\mu\mathrm{s}$ corresponding to T_{sim}. The MSD curves show TAD and reproduce both the transient and asymptotic regime. (**Bottom**) Positional PDFs are computed from the C.o.M diffusive dynamics with initial condition $\mathbf{X}_0 = 0$ for both ML-GLE and MD reference dynamics.

and their asymptotic behaviour computationally expensive, even with an already CG system.
The neural auto-regressive generative model we presented is capable of learning the *coarse-grained* time integrator for slower normal modes by means of a maximum likelihood training. Thus, for each of these processes the long-term behaviour can be extrapolated, providing a way to simulate single polymer degrees of freedom and implicitly learn the tail of the kernel function, accelerating consequently the observation of the diffusive regime and estimation of the diffusion coefficient with a remarkable gain in computational time, with respect to performing the full-size CG MD simulation.

A few key points need to be addressed at this point. First, training data comes from a CG system, which implies an upper bound on the method accuracy; this is determined by the quality of the CG approximation as compared to the *All-Atom* simulation. Secondly, although the bijective mapping from monomer coordinates to normal modes allows for configuration reconstruction, the statistical independence approximation could fail to reproduce topologically forbidden self-intersections in real space, implying that correlations may be indeed present especially for higher frequency modes, corresponding to local configurations, and thus affected by volume exclusion effects.
In addition, as discussed above, the generative model for modes dynamics deviates from Gaussian behaviour in the PDFs tails, making large fluctuations more likely but unrealistic. A possible improvement in this direction could involve employing architecture that are natively able to capture correlations in 3D sequences, such as RNN or neural networks with attention mechanisms, like Transformers [63], instead of a simple MLP. These more advanced architectures could be further enhance the generalization capabilities over *unobserved* thermodynamic coordinates, chemical space, or system size, when trained on a large dataset of trajectories coming from the same physical system, but at different conditions. For example, training could be performed on simulation data at different temperatures, that would act as conditioning static features.

Future research perspectives also include developing methods for generating larger systems reproducing simul-

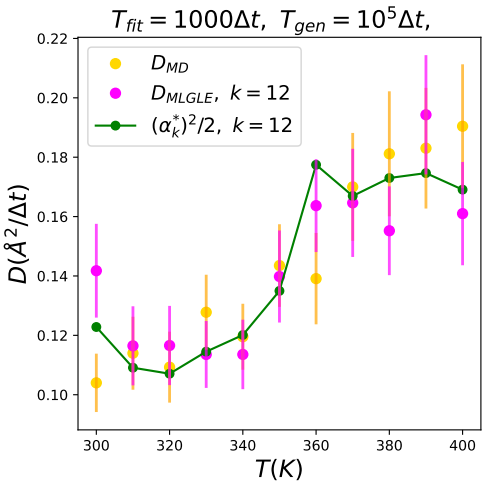

Fig. 14: Diffusion coefficients comparison for 11 temperatures, between 300 K and 400 K. In *yellow* the reference MD diffusion coefficients, D_{MD} and in *magenta* the coefficients as estimated from the generated C.o.M dynamics with ML-GLE with $k = 12$ auto-regressive generated modes. In *green* we also provide the diffusion coefficient obtained from the GLE solution fitting, $D_{GLE} = (\alpha_k^*)^2/2$. confidence intervals are computed as in Fig. 10 for the MD diffusion coefficient estimate.

taneously structural and dynamical properties, using the dynamics generated by the ML-GLE to back-map to the original *coarse-grained* description without loosing the dynamical correlations. The ML-GLE framework we propose could serve therefore as a starting point for more complicated machine learning methods, using the proposed dynamics as a first approximation but aiming at the reconstruction of the full-size fine-grained system, therefore accelerating the full-size collective polymer simulation.

Another important research direction revolves around how this framework could be adapted to different systems, from biological to artificial, including methods for extracting slow varying degrees of freedom from complex molecules in order to describe its long term dynamics. In this respective, anomalous diffusion has been discovered in numerous systems, mostly in disordered media [24].
Some interesting examples include external tracer particles in biological cells [64, 65], and in artificially crowded systems [21, 66]. Protein simulations studies have reported anomalous sub-diffusive behaviour as well [67, 68, 69]. All these systems exhibit transient behaviour and therefore this approach could pave the way for data-driven modeling of diffusing macro-molecules within the GLE framework, both from *in vitro* and *in vivo* data sources [70].

Acknowledgments

This work was performed at SimatLab, A joint public-private laboratory dedicated to the modeling of polymeric materials. This laboratory is supported by Michelin, Clermont-Auvergne University (UCA), CHU Clermont and CNRS. SimatLab acknowledges support received from the *Agence Nationale de la Recherche* of the French government through the program "Investissements d'Avenir" (grant no. 16-IDEX-0001 CAP 20-25).
We are grateful to the Mésocentre Clermont-Auvergne for providing help, computing and storage resources. All Computations have been performed on the supercomputer facilities of the Mésocentre Clermont-Auvergne of the Université Clermont Auvergne.

References

[1] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686–707.
[2] X. Meng, Z. Li, D. Zhang, G. E. Karniadakis, Ppinn: Parareal physics-informed neural network for time-dependent pdes, Computer Methods in Applied Mechanics and Engineering 370 (2020) 113250.
[3] N. B. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. M. Stuart, A. Anandkumar, Neural operator: Learning maps between function spaces with applications to pdes., J. Mach. Learn. Res. 24 (2023) 1–97.

[4] A. Krishnapriyan, Characterizing possible failure modes in physics-informed neural networks (characterizing-pinns-failure-modes) v0. 1.0, Technical Report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2021.

[5] J. Fu, D. Xiao, R. Fu, C. Li, C. Zhu, R. Arcucci, I. M. Navon, Physics-data combined machine learning for parametric reduced-order modelling of nonlinear dynamical systems in small-data regimes, Computer Methods in Applied Mechanics and Engineering 404 (2023) 115771.

[6] P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, Journal of Fluid Mechanics 656 (2010) 5–28.

[7] S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences 113 (2016) 3932–3937.

[8] S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations, Science Advances 3 (2017) e1602614.

[9] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 473 (2017) 20160446.

[10] H. Mori, Transport, collective motion, and brownian motion, 1965. URL: https://academic.oup.com/ptp/article-lookup/doi/10.1143/PTP.33.423. doi:10.1143/PTP.33.423.

[11] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, Schnet – a deep learning architecture for molecules and materials, The Journal of Chemical Physics 148 (2018) 241722.

[12] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, C. Clementi, Machine learning of coarse-grained molecular dynamics force fields, ACS Central Science 5 (2019) 755–767.

[13] W. Wang, R. Gómez-Bombarelli, Coarse-graining auto-encoders for molecular dynamics, npj Computational Materials 5 (2019) 125.

[14] Z. Shireen, H. Weeratunge, A. Menzel, A. W. Phillips, R. G. Larson, K. Smith-Miles, E. Hajizadeh, A machine learning enabled hybrid optimization framework for efficient coarse-graining of a model polymer, npj Computational Materials 8 (2022) 224.

[15] J. Bakarji, D. M. Tartakovsky, Data-driven discovery of coarse-grained equations, Journal of Computational Physics 434 (2021) 110219.

[16] R. Zwanzig, M. Bixon, Hydrodynamic theory of the velocity correlation function, Physical Review A 2 (1970) 2005–2012.

[17] Z. Li, H. S. Lee, E. Darve, G. E. Karniadakis, Computing the non-markovian coarse-grained interactions derived from the mori–zwanzig formalism in molecular systems: Application to polymer melts, The Journal of Chemical Physics 146 (2017) 014104.

[18] Z. Li, X. Bian, X. Li, G. E. Karniadakis, Incorporation of memory effects in coarse-grained modeling via the mori-zwanzig formalism, The Journal of Chemical Physics 143 (2015) 243128.

[19] V. Klippenstein, M. Tripathy, G. Jung, F. Schmid, N. F. A. Van Der Vegt, Introducing memory in coarse-grained molecular simulations, The Journal of Physical Chemistry B 125 (2021) 4931–4954.

[20] R. Zwanzig, Nonlinear generalized langevin equations, Journal of Statistical Physics 9 (1973) 215–220.

[21] K. Norregaard, R. Metzler, C. M. Ritter, K. Berg-Sørensen, L. B. Oddershede, Manipulation and motion of organelles and single molecules in living cells, Chemical Reviews 117 (2017) 4342–4375.

[22] F. Höfling, T. Franosch, Anomalous transport in the crowded world of biological cells, Reports on Progress in Physics 76 (2013) 046602.

[23] J.-H. Jeon, N. Leijnse, L. B. Oddershede, R. Metzler, Anomalous diffusion and power-law relaxation of the time averaged mean squared displacement in worm-like micellar solutions, New Journal of Physics 15 (2013) 045011.

[24] J.-P. Bouchaud, A. Georges, Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications, Physics Reports 195 (1990) 127–293.

[25] A. J. Spakowitz, Transient anomalous diffusion in a heterogeneous environment, Frontiers in Physics 7 (2019) 119.

[26] D. Molina-Garcia, T. Sandev, H. Safdari, G. Pagnini, A. Chechkin, R. Metzler, Crossover from anomalous to normal diffusion: truncated power-law noise correlations and applications to dynamics in lipid bilayers, New Journal of Physics 20 (2018) 103027.

[27] S. A. McKinley, L. Yao, M. G. Forest, Transient anomalous diffusion of tracer particles in soft matter, Journal of Rheology 53 (2009) 1487–1506.

[28] F. Grogan, H. Lei, X. Li, N. A. Baker, Data-driven molecular modeling with the generalized langevin equation, Journal of Computational Physics 418 (2020) 109633.

[29] C. Ma, J. Wang, W. E, Model reduction with memory and the machine learning of dynamical systems, Commun. Comput. Phys. 25 (2019). ArXiv:1808.04258 [physics, stat].

[30] K. Lee, K. T. Carlberg, Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders, Journal of Computational Physics 404 (2020) 108973.

[31] A. Chorin, P. Stinis, Problem reduction, renormalization, and memory, Communications in Applied Mathematics and Computational Science 1 (2006) 1–27.

[32] M. Berkowitz, J. D. Morgan, J. A. McCammon, Generalized langevin dynamics simulations with arbitrary time-dependent memory kernels, The Journal of Chemical Physics 78 (1983) 3256–3261.

[33] J. Fricks, L. Yao, T. C. Elston, M. G. Forest, Time-domain methods for diffusive transport in soft matter, SIAM Journal on Applied Mathematics 69 (2009) 1277–1308.

[34] A. Davtyan, J. F. Dama, G. A. Voth, H. C. Andersen, Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence, The Journal of Chemical Physics 142 (2015) 154104.

[35] M. Ceriotti, G. Bussi, M. Parrinello, Langevin equation with colored noise for constant-temperature molecular dynamics simulations, Physical Review Letters 102 (2009) 020601.

[36] M. Ceriotti, M. Parrinello, T. E. Markland, D. E. Manolopoulos, Efficient stochastic thermostatting of path integral molecular dynamics, The Journal of Chemical Physics 133 (2010) 124104.

[37] N. Schaudinnus, B. Bastian, R. Hegger, G. Stock, Multidimensional langevin modeling of nonoverdamped dynamics, Physical Review Letters 115 (2015) 050602.

[38] S. Cao, Y. Qiu, M. L. Kalin, X. Huang, Integrative generalized master equation: A method to study long-timescale biomolecular dynamics via the integrals of memory kernels, The Journal of Chemical Physics 159 (2023) 134106.

[39] A. J. Chorin, F. Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, Proceedings of the National Academy of Sciences 112 (2015) 9804–9809.

[40] P. Xie, R. Car, W. E, Ab initio generalized langevin equations, ArXiv (2022).

[41] K. Endo, K. Tomobe, K. Yasuoka, Multi-step time series generator for molecular dynamics, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018).
[42] Y. Zhu, Y.-H. Tang, C. Kim, Learning stochastic dynamics with statistics-informed neural network, Journal of Computational Physics 474 (2023) 111819.
[43] R. Kubo, The fluctuation-dissipation theorem, Reports on Progress in Physics 29 (1966) 255–284.
[44] R. Morgado, F. A. Oliveira, G. G. Batrouni, A. Hansen, Relation between anomalous and normal diffusion in systems with memory, Physical Review Letters 89 (2002) 100601.
[45] S. A. McKinley, H. D. Nguyen, Anomalous diffusion and the generalized langevin equation, SIAM Journal on Mathematical Analysis 50 (2018) 5119–5160.
[46] M. Doi, S. F. Edwards, The theory of polymer dynamics, volume 27, Oxford Science Publications, 1989. doi:10.1002/pol.1989.140270706.
[47] I. Goychuk, Viscoelastic Subdiffusion: Generalized Langevin Equation Approach, volume 150, 1 ed., Wiley, 2012, p. 187–253. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118197714.ch5. doi:10.1002/9781118197714.ch5.
[48] R. M., R. H. Colby, Polymer Physics, 2003.
[49] F. Gilani, D. Giannakis, J. Harlim, Kernel-based prediction of non-markovian time series, Physica D: Nonlinear Phenomena 418 (2021) 132829.
[50] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems 2 (1989) 303–314.
[51] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nature Communications 13 (2022) 2453.
[52] V. Garcia Satorras, E. Hoogeboom, M. Welling, E (n) equivariant graph neural networks, arXiv e-prints (2021) arXiv–2102.
[53] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, P. Battaglia, Learning to simulate complex physics with graph networks, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 8459–8468. URL: https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.
[54] C. Yang, W. Gao, D. Wu, C. Wang, Learning to simulate unseen physical systems with graph neural networks, arXiv preprint arXiv:2201.11976 (2022).
[55] A. Dequidt, J. G. Solano Canchaya, Bayesian parametrization of coarse-grain dissipative dynamics models, The Journal of Chemical Physics 143 (2015).
[56] G. Williams, D. C. Watts, Non-symmetrical dielectric relaxation behaviour arising from a simple empirical decay function, Transactions of the Faraday Society 66 (1970) 80.
[57] K. M. Bennett, K. M. Schmainda, R. Bennett (Tong), D. B. Rowe, H. Lu, J. S. Hyde, Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model, Magnetic Resonance in Medicine 50 (2003) 727–734.
[58] S. I. Simdyankin, N. Mousseau, Relationship between dynamical heterogeneities and stretched exponential relaxation, Physical Review E 68 (2003) 041110.
[59] J. C. Phillips, Stretched exponential relaxation in molecular and electronic glasses, Reports on Progress in Physics 59 (1996) 1133–1207.
[60] J. Zhong, R. B. Diener, D. A. Steck, W. H. Oskay, M. G. Raizen, E. W. Plummer, Z. Zhang, Q. Niu, Shape of the quantum diffusion front, Physical Review Letters 86 (2001) 2485–2489.
[61] G. Mélard, R. Roy, On confidence intervals and tests for autocorrelations, Computational Statistics & Data Analysis 5 (1987) 31–44.
[62] V. Calandrini, E. Pellegrini, P. Calligari, K. Hinsen, G. Kneller, nmoldyn - interfacing spectroscopic experiments, molecular dynamics simulations and models for time correlation functions, École thématique de la Société Française de la Neutronique 12 (2011) 201–232.
[63] J. Tong, L. Xie, K. Zhang, Probabilistic decomposition transformer for time series forecasting, in: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), SIAM, 2023, pp. 478–486.
[64] I. Bronstein, Y. Israel, E. Kepten, S. Mai, Y. Shav-Tal, E. Barkai, Y. Garini, Transient anomalous diffusion of telomeres in the nucleus of mammalian cells, Physical Review Letters 103 (2009) 018102.
[65] G. Seisenberger, M. U. Ried, T. Endreß, H. Büning, M. Hallek, C. Bräuchle, Real-time single-molecule imaging of the infection pathway of an adeno-associated virus, Science 294 (2001) 1929–1932.
[66] J. Szymanski, M. Weiss, Elucidating the origin of anomalous diffusion in crowded fluids, Physical Review Letters 103 (2009) 038102.
[67] M. Javanainen, H. Hammaren, L. Monticelli, J.-H. Jeon, M. S. Miettinen, H. Martinez-Seara, R. Metzler, I. Vattulainen, Anomalous and normal diffusion of proteins and lipids in crowded lipid membranes, Faraday Discuss. 161 (2013) 397–417.
[68] K. Chen, B. Wang, S. Granick, Memoryless self-reinforcing directionality in endosomal active transport within living cells, Nature Materials 14 (2015) 589–593.
[69] J.-H. Jeon, M. Javanainen, H. Martinez-Seara, R. Metzler, I. Vattulainen, Protein crowding in lipid bilayers gives rise to non-gaussian anomalous lateral diffusion of phospholipids and proteins, Physical Review X 6 (2016) 021006.
[70] J.-H. Jeon, V. Tejedor, S. Burov, E. Barkai, C. Selhuber-Unkel, K. Berg-Sørensen, L. Oddershede, R. Metzler, In vivo anomalous diffusion and weak ergodicity breaking of lipid granules, Physical Review Letters 106 (2011) 048103.

Appendix A GLE solution with the Prony series OU processes

We report here the GLE solution steps as appear in [27]. Starting from a 1D GLE in the form,

$$M\ddot{X}(t) = -\int_0^t \Gamma(t-s)\dot{X}(s)ds + F(t)$$

and having specified the noise term and the kernel function as a set of *Ornstein-Uhlenbeck* processes and *Prony series* respectively, applying the Laplace transform to all terms.
The solution in Laplace domain reads,

$$\tilde{X}(z) = \frac{z^{-1}\tilde{F}(z)}{Mz + \tilde{\Gamma}(z)}. \tag{37}$$

where the noise's transform can be written as,

$$\tilde{F}(z) = \sum_{j=1}^{N} \tilde{F}_j(z) = \sum_{j=1}^{N} \sqrt{2k_B T\eta\gamma_j}\mathcal{L}\Big\{\int_0^t e^{-\gamma_j(t-s)}dW_j(s)\Big\}(z)$$

By computing the Laplace transform of the OU integral solution, the numerator becomes,

$$z^{-1}\tilde{F}(z) = \sqrt{2k_B T\eta}\sum_{k=1}^{N}\frac{\sqrt{\gamma_k}}{z+\gamma_k}\mathcal{L}\{W_k(t)\}(z)$$

Denoting $\sigma^2 = 2k_B T\eta^{-1}$, and $\tilde{W}_j(z) = \mathcal{L}\{W_j(t)\}(z)$ as the Wiener process transform, the solution in zero-mass limit becomes,

$$\tilde{X}(z) = \frac{z^{-1}\tilde{F}(z)}{\tilde{\Gamma}(z)} = \sigma\frac{\sum_{j=1}^{N}\frac{\sqrt{\gamma_j}}{z+\gamma_j}\tilde{W}_j(t)(z)}{\sum_{j=1}^{N}\frac{1}{z+\gamma_j}}.$$

Further denoting as $p(z) = \prod_{j=1}^{N}(z+\gamma_j)$ one can write the solution in a more compact form in the following way,

$$\tilde{X}(z) = \sigma\sum_{j=1}^{N}\frac{p(z)}{p'(z)(z+\gamma_j)}\sqrt{\gamma_j}\tilde{W}_j(z).$$

The key point for solving the equation is recognizing that $p(z)/p'(z+\alpha_k)$ is a rational function with a numerator and denominator which have the same degree. The simple pole of this function are the zeros of $p'(z)$, $\{-r_j\}_{j=1}^{N}$.
Hence, the following decomposition holds,

$$\frac{p(z)}{p'(z)(z+\gamma_k)} = \frac{1}{N} + \sum_{k=1}^{N-1}\frac{p(-r_j)}{p''(-r_j)(-r_j+\gamma_k)}\frac{1}{z+r_j}$$

In Laplace space, integration by part equals

$$\frac{1}{z+r_j} = \frac{1}{r_j}(1 - \frac{z}{z+r_j}),$$

and one has the following result,

$$\tilde{X}(z) = \sigma(S_1 - S_2)$$

where,

$$S_1 = \sum_{j=1}^{N}\left(\frac{1}{N} + \sum_{l=1}^{N-1}\frac{p(-r_l)}{p''(-r_l)(\gamma_j - r_l)}\frac{1}{r_l}\right)\sqrt{\gamma_j}\tilde{W}_j(z)$$

$$S_2 = \sum_{j=1}^{N}\sum_{l=1}^{N-1}\frac{p(-r_l)}{p''(r_l)(\gamma_j - r_l)r_l}\frac{z}{z+r_l}\sqrt{\gamma_j}\tilde{W}_j(z)$$

The solution can be decomposed in a frequency independent part and a frequency dependent part. As we will see, these terms correspond in the time domain, to the a simple Brownian motion and a term involving sum of more complicated processes, respectively.
To see this, one can simplify further S_1,

$$S_1 = \frac{p(0)}{p'(0)} \sum_{j=1}^{N} \frac{1}{\sqrt{\gamma_j}} \tilde{W}_j(z)$$

Inverting back to time domain, we can use the properties of the Brownian motion,

$$\sum_{j=1}^{N} \frac{1}{\sqrt{\gamma_j}} W_j(t) = \left(\sum_{j=1}^{N} \frac{1}{\sqrt{\gamma_j}} \right)^{1/2} B(t) = \frac{1}{\sqrt{N\bar{\tau}}} B(t)$$

Concerning S_2, it can be written as,

$$S_2 = \sum_{l=1}^{N-1} \frac{p(-r_l)}{p''(-r_l) r_l} \frac{z}{z + r_l} \sum_{k=1}^{N} \frac{\sqrt{\gamma_k}}{\gamma_k - r_l} \tilde{W}_j(z) = \sum_{l=1}^{N-1} \frac{p(-r_l)}{p''(-r_l) r_l} \frac{z}{z + r_l} \tilde{\xi}_l(z)$$

where $\tilde{\xi}_l(z) = \sum_{j=1}^{N} \frac{\sqrt{\gamma_j}}{\gamma_j - r_l} \tilde{W}_j(z), \forall l \in \{1, ..., N\}$ is set of statistically independent Brownian motion processes whose scale parameter depends on the whole diffusive spectrum, providing collective information about the dynamics.

Inverting back in the time domain,

$$\sum_{l=1}^{N-1} \frac{p(-r_l)}{p''(-r_l) r_l} \int_0^t dz \frac{z}{z + r_l} \tilde{\xi}_l(z) e^{zt} = \sum_{l=1}^{N-1} \frac{p(-r_l)}{p''(-r_l) r_l} \left(\sum_{j=1}^{N} \frac{\gamma_j}{(\gamma_j - r_l)^2} \right)^{1/2} Z_l.$$

where $\{Z_l\}_{l=1}^{N-1}$ is a set of independent OU processes with parameters corresponding to the roots of $p'(z)$, $\{-r_l\}_{l=1}^{N-1}$.

Wrapping up the two inverted terms in time domain, one can write the final path solution,

$$X(t) = \frac{\sigma}{\sqrt{N\bar{\tau}}} B(t) + \sigma \sum_{l=1}^{N-1} c_j Z_l.$$

The set of coefficients $\{c_l\}_{l=1}^{N-1}$ are such that,

$$c_l = \frac{p(-r_l)}{p''(-r_l) r_l} \left(\sum_{j=1}^{N} \frac{\gamma_j}{(\gamma_j - r_l)^2} \right)^{1/2}.$$

Appendix B Gaussian PDFs

Since the Neural Autoregressive (NAR) Model is optimized over a Multivariate Normal distribution modeling the process Gaussian increments $\Delta\mathbf{z}_j$, we demonstrate in Fig. 15 that this is a suitable choice as both the original and generated process adhere to this hypothesis, as shown by the empirical and fit Normal PDFs.

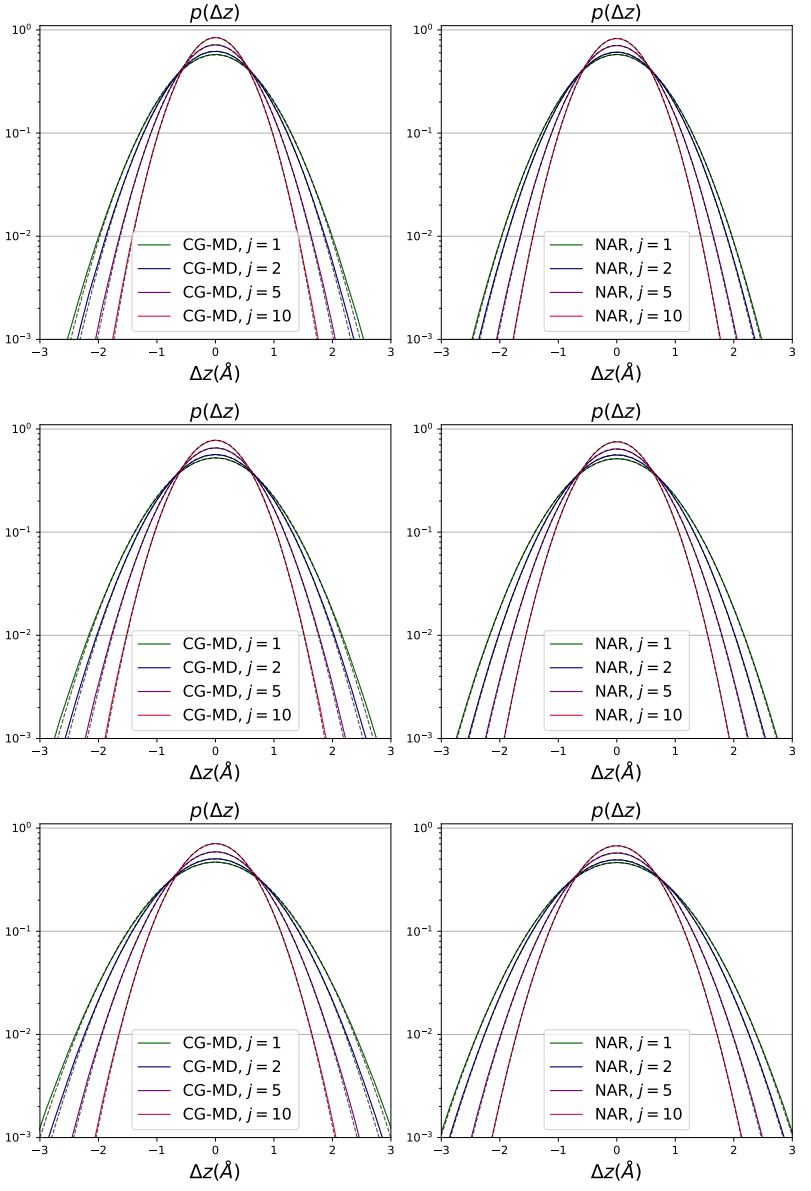

Fig. 15: PDFs of $\Delta\mathbf{z}$ averaged over the 3 spatial dimensions for $T = 300\,\mathrm{K}$ (**Top**), $T = 340\,\mathrm{K}$ (**Center**), $T = 400\,\mathrm{K}$ (**Bottom**). The fit Gaussian PDFs are obtained with parameters computed from the CG-MD simulation (**left-side**) and the ML-GLE generated processes (**right-side**).

The same goes for the Gaussian increments related the C.o.M fluctuations $\Delta\mathbf{X}$ against the ML-GLE model $\Delta\mathbf{X}_k$. In Fig. 16, we can see that both the MD and generated process follow the characteristic Normal PDF, even if small deviations from Gaussianity are noticeable on the CG-MD simulation PDFs tails corresponding to regions covered with small probability ($< 10^{-3}$). Furthermore, the CG-MD and the ML-GLE increments mismatch in their PDFs tails, which is a result of the GLE solution approximation with k modes.

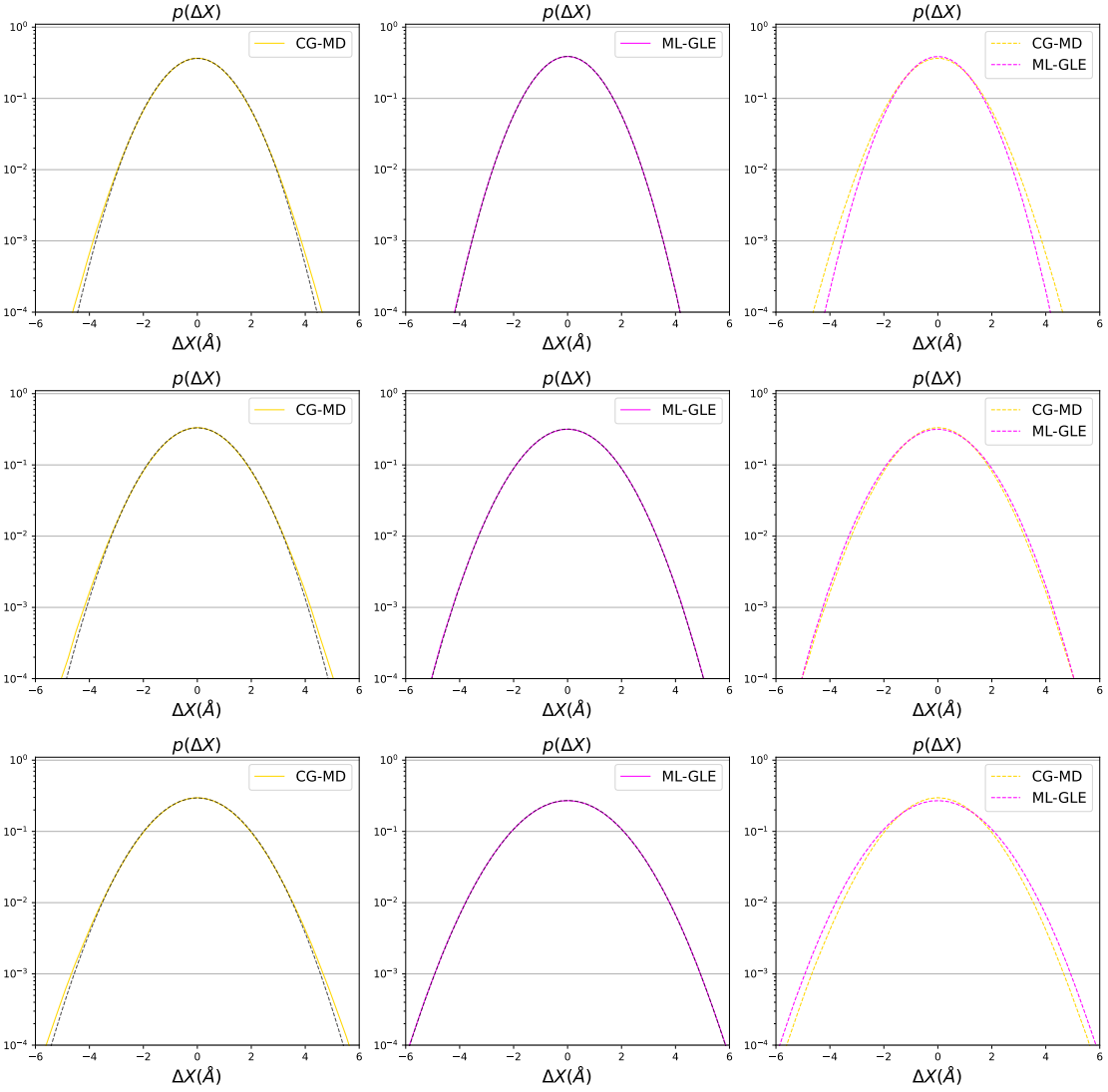

Fig. 16: PDFs of $\Delta\mathbf{z}$ averaged over the 3 spatial dimensions for $T = 300\,\mathrm{K}$ (**Top**), $T = 340\,\mathrm{K}$ (**Center**), $T = 400\,\mathrm{K}$ (**Bottom**). The comparison between the CG-MD and ML-GLE process PDFs are shown on the right side.

Appendix C Loss vs Batchsize

We show here training results in order to investigate its sensibility with respect to batchsize. We trained the same NAR network with the same amount of data, hence using simulations up to $T_{train} = 1000\Delta t$ and historical window $m = 128$. Validation and training curves for 100 epochs are represented in Fig. 17. The different trainings were executed with the same random seed which enforces the same initial weights on the network.
One can clearly notice that changing batchsize does not influence the validation minima in an appreciable way. However, at increasing batchsize, an observable tendency to reduce the validation train gap can be noticed. In terms of generation quality, we chose the best batchsize at 128 by trial and error, since training with small batches did not provide a satisfying result.

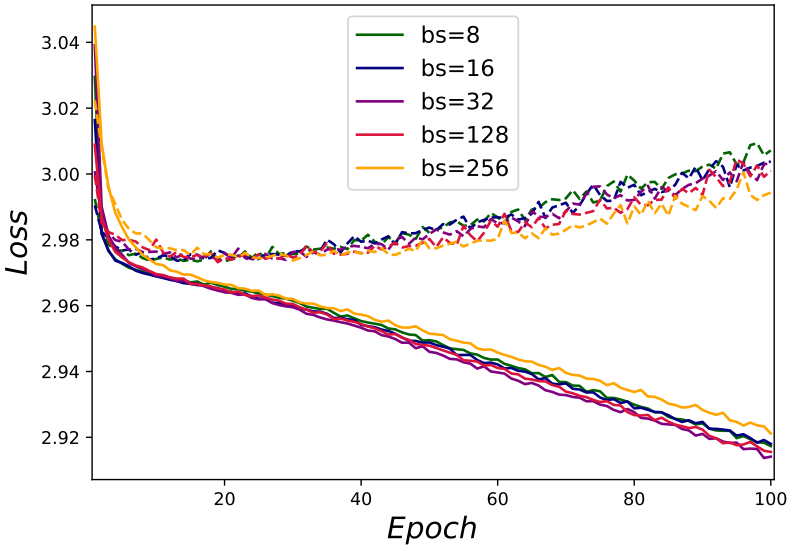

Fig. 17: Training (*solid*) and validation (*dashed*) loss for 100 epochs at $8, 32, 64, 128$ and 256 batchsize.

Appendix D Normal modes probability plots

In Fig. 18 we show normal probability plots associated with the normal modes PDFs as generated form the NAR generative model compared to the reference ones obtained from the CG system MD simulations.
We remark how the effect of temperature increase on deviations from normality is equivalent to considering faster modes. In other words, the NN is more capable of reproducing accurately the normal modes PDFs when the input trajectory is longer (i.e. higher order memory). When temperature decreases, the slowest mode becomes slower and we can observe a small, but more pronounced non-Gaussian behaviour, as well as a higher discrepancy on the PDF mean and scale parameters.

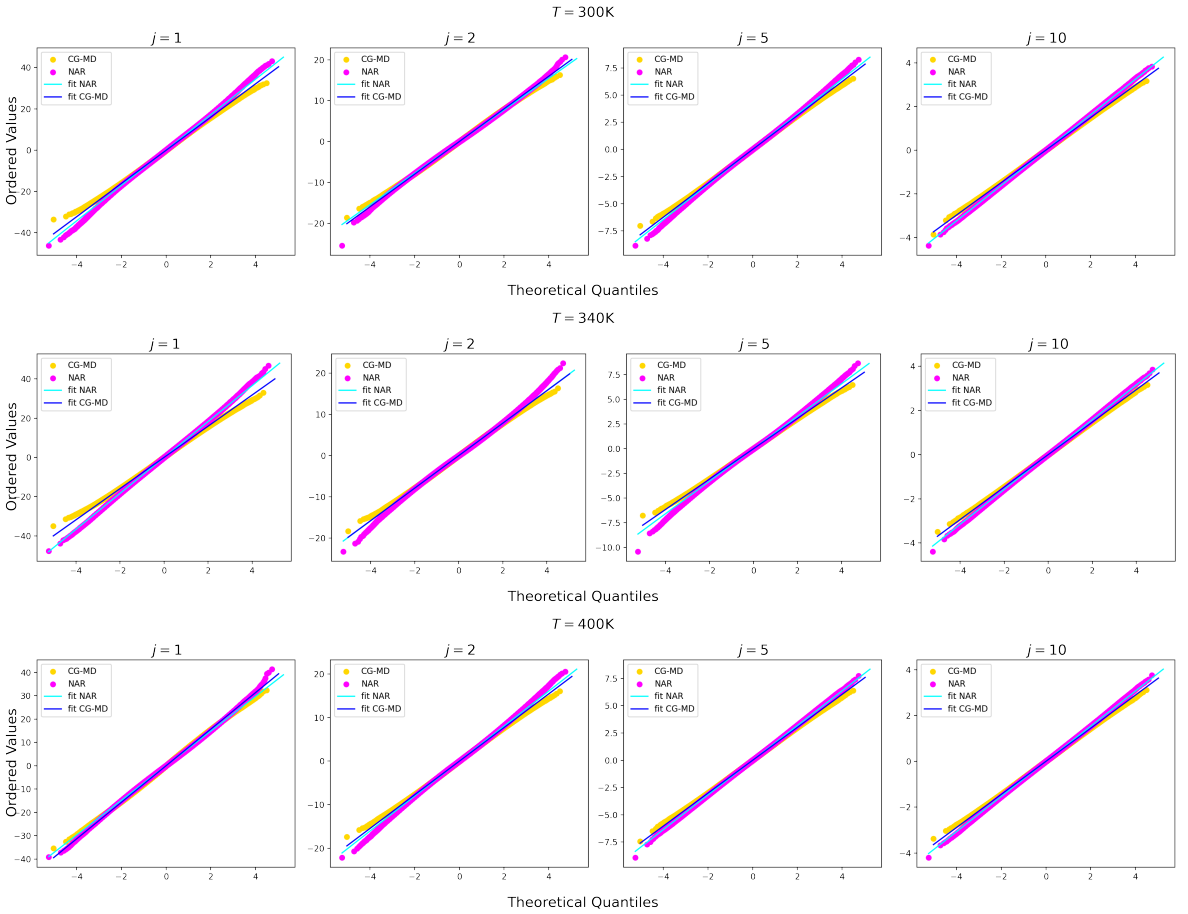

Fig. 18: Normal Probability plots for $T = 300\,\mathrm{K}$ (**Top**), $T = 340\,\mathrm{K}$ (**Center**), $T = 400\,\mathrm{K}$ (**Bottom**). The normal modes PDFs unscaled quantiles are shown against theoretical standard PDF quantiles for the 1*st*, 2*nd*, 5*th* and 10*th* mode. Slower modes (in *magenta*) tend to exhibit small departures from normality on the extreme values, with respect to the MD ones (in *gold*). This effect tends instead to disappear for faster modes. In *cyan* and *blue* are shown the NAR and MD PDF fit respectively, which depend on the estimated mean and scale parameters.