The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 05 2023

Combining machine learning and molecular simulations to predict the stability of amorphous drugs

Special Collection: Machine Learning Hits Molecular Simulations

Trent Barnard ; Gabriele C. Sosso

J. Chem. Phys. 159, 014503 (2023)
https://doi.org/10.1063/5.0156222

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Combining machine learning and molecular simulations to predict the stability of amorphous drugs

Cite as: J. Chem. Phys. 159, 014503 (2023); doi: 10.1063/5.0156222
Submitted: 27 April 2023 • Accepted: 8 June 2023 •
Published Online: 5 July 2023

Trent Barnard and Gabriele C. Sosso[a)]

AFFILIATIONS

Department of Chemistry, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a)]**Author to whom correspondence should be addressed:** g.sosso@warwick.ac.uk

ABSTRACT

Amorphous drugs represent an intriguing option to bypass the low solubility of many crystalline formulations of pharmaceuticals. The physical stability of the amorphous phase with respect to the crystal is crucial to bring amorphous formulations into the market—however, predicting the timescale involved with the onset of crystallization *a priori* is a formidably challenging task. Machine learning can help in this context by crafting models capable of predicting the physical stability of any given amorphous drug. In this work, we leverage the outcomes of molecular dynamics simulations to further the state-of-the-art. In particular, we devise, compute, and use "solid state" descriptors that capture the dynamical properties of the amorphous phases, thus complementing the picture offered by the "traditional," "one-molecule" descriptors used in most quantitative structure–activity relationship models. The results in terms of accuracy are very encouraging and demonstrate the added value of using molecular simulations as a tool to enrich the traditional machine learning paradigm for drug design and discovery.

I. INTRODUCTION

Most modern pharmaceutical drugs are packaged as crystalline formulations.[1] The crystalline structure has significant effects on several physical properties of the drug, such as its solubility, its stability, and its bioavailability.[2] Crucially, almost 90% of pharmaceutical drugs are categorized as poorly water soluble,[3,4] which clearly limits their effectiveness, chiefly in terms of bioavailability.

Packaging pharmaceutical drugs as amorphous formulations represents a viable way forward in order to improve the solubility of modern drug formulations[5] as they present several benefits in comparison to crystalline drugs. First, most amorphous compounds are intrinsically much more soluble than their crystalline counterparts.[6–8] As such, amorphous drugs typically act more quickly than crystalline drugs.[9,10] In addition, amorphous drugs can be more easily packaged into different formulations—such as tablets, capsules, or suspensions.[8,11] In fact, the lack of crystalline structure can also allow for greater flexibility in designing drug delivery systems with specific properties, such as sustained release or targeted delivery.[8]

While amorphous drugs appear to have an edge over their crystalline counterparts, they also have some disadvantages that can make their development and formulation challenging—chiefly their lack of stability. Amorphous solids are almost always metastable with respect to their crystalline phases, which means that amorphous drugs have a tendency to crystallize[12]—within a timescale that is very challenging to predict. This represents a serious problem,[12] in that the properties of the crystalline form might differ from that of the amorphous phase—which poses a severe clinical risk. In addition, the structural relaxation of the glass alone might alter the functional properties of the amorphous formulation.[13] It is also important to note that the production of amorphous drugs can be more challenging than that of crystalline drugs, requiring specialized manufacturing techniques.[14] When it comes to delaying the onset of the crystallization process, for the purposes of prolonging the shelf life of the amorphous drug formulation, the usage of so-called amorphous solid dispersions (ASDs) represents a common strategy.[15] An ASD is an heterogeneous system obtained by incorporating the amorphous phase within a (usually) polymeric matrix. This helps to stabilize the amorphous phase, thus delaying the onset of

crystallization. ASDs can be created using various different techniques, including spray drying and co-precipitation.[16,17]

For the purposes of this work, it is important to mention the actual process by which most amorphous drugs formulations are actually obtained experimentally, namely, by quenching from the melt. This process involves heating the drug to a temperature higher than its melting point (T_m) so that it forms a liquid. This liquid is then rapidly cooled by immersing it in liquid nitrogen, placing it on a cold surface or blowing cold gas over it. If the drug is cooled rapidly enough, it is prevented from reorganizing itself into its crystallize form, and an amorphous solid is formed. To create an ASD, the exact same process is used—except the amorphous drug is melted together with a polymer. This technique is simple and effective and can be used with a wide range of drugs and polymers. However, the high temperatures used can lead to drug degradation.[18] Other approaches to manufacture amorphous formulations can be used, such as spray drying[16] or freeze drying. The experimental data generated for use in this work,[19,20] however, refer to amorphous drugs obtained via melt-quenching.

One of the most pressing issues, in terms of adopting amorphous drugs as reliable formulations for modern pharmaceuticals, is that it is very challenging to predict *a priori* their physical stability. In addition, measuring the physical stability experimentally can be a challenging endeavor in itself, given that the timescales involved with the onset of the crystallization process range from seconds to years—depending on the specific amorphous drug. As such, building a computational framework capable to predict the physical stability of novel drugs would be massively beneficial for the pharmaceutical industry. Indeed, a few attempts to harness machine learning (ML) algorithms as a tool to predict the physical stability of amorphous drugs can already be found within the recent literature.[21–23] As it often is the case, the main hurdle in this context is the rather limited amount of experimental data available to us—an issue that is particularly stark in the case of the physical stability of amorphous drugs. To complicate matters even further, the stability of an amorphous drug is affected by many different external variables, such as the storage conditions (chiefly temperature and humidity) as well as the method used to create the amorphous drug in the first place. Clearly, these factors constitute a major source of variability and uncertainty within the experimental data, which is especially challenging to pinpoint in the absence of exhaustive information about both the manufacturing and the storage of the amorphous formulation.

In this work, we improve on the current state-of-the-art in terms of predicting the stability of amorphous drugs. This is achieved by adopting a combination of methodological improvements in terms of the underlying machine learning algorithms and, crucially, by leveraging the outcomes of molecular dynamics simulations to both complement and enhance the portfolio of the descriptors (or features, or fingerprints) traditionally used in the context of ML for drug design and/or discovery.

Due to the extremely limited amount of reliable data on the stability of amorphous drugs, we focus on two key properties that are closely related to the stability of amorphous pharmaceuticals. First, we consider the glass transition temperature (T_g), which is correlated with the propensity of the system to form a disordered solid as opposed to a crystal in the first place[24] and it also correlates with a good extent with the physical stability of the amorphous phase.[25] Amorphous drugs characterized by high values of T_g are typically more stable and thus less likely to crystallize within a given timescale.[26] The second property we focus on in this work is the so-called "crystallization class." This is a classification system originally developed by Baird *et al.*,[27] whereby drug-like molecules can be separated into three distinct classes, based on their propensity to crystallize during a specific annealing (heat/cool/heat) cycle. These three classes are defined as follows:

- **Class I** drugs. Crystallization is observed while cooling from the melt. Clearly, drugs that belong to this class are not suitable for amorphous drug formulations.
- **Class II** drugs. No crystallization is observed when cooling from the melt to below T_g. However, the system will crystallize when re-heated above T_g during the annealing cycle. This is perhaps the most interesting class of amorphous drugs from a fundamental perspective.
- **Class III** drugs. No crystallization is observed, either during the quenching from the melt or the annealing cycle. Drugs belonging to this class are probably the most suitable candidates from a practical point of view in terms of amorphous drugs formulations.

The state-of-the art, with respect to machine learning-based models for predicting both the T_g and the crystallization class of amorphous drugs, is largely defined by the pioneering work of Alhalaweh *et al.*, particularly Refs. 19 and 28. Indeed, our work relies on the data that were made available via these publications.

This paper is organized as follows: we begin with Sec. II, discussing the data in our possession, as well as the details of the MD simulations we have used to generate the models of amorphous drugs. We also discuss the machine learning algorithms we have used, with specific reference to optimization strategies. Novel aspects of our work in this context include the usage of genetic algorithms (GAs) to improve the accuracy of specific descriptors, as well as the combination of multiple classes of descriptors via ensemble learning.

In Sec. III, we present the outcomes of our models in terms of the prediction of T_g (a regression problem) and of the crystalline class (a classification problem) as well. We start with the results pertaining to isolated classes of different descriptors, all of them obtained by considering single molecules in isolation, and discuss the accuracy improvements obtained via both descriptor optimization and ensemble learning. We then move to the descriptors we have obtained by leveraging actual MD simulations. To be specific, we have generated amorphous models for each drug within our dataset. As a striking result in itself, the T_g obtained via MD simulations correlates very well with the experimental results. This serves as a validation of our computational protocol, and it strengthens the notion that MD simulations can be used to extract information about amorphous systems that are not directly available when considering single molecules "in vacuum." From our MD simulations, we have computed both the diffusion coefficient and the relaxation time of the supercooled liquid phase at a specific temperature that varies for different drugs according to the calculated T_g. These dynamical properties can be used as descriptors themselves—and indeed, we show that they make important contributions to the accuracy of our ML models. Finally, we combine these "solid state" descriptors to the "one-molecule" descriptors discussed in

Sec. II to yield a ML model that, we argue, represents a significant improvement with respect to the state-of-the-art.

It is crucial at this stage to remind the reader that this field is in its very infancy: with the very limited amount of experimental data in our possession, we are in no position to argue that our model can be reliably used to predict the physical stability of amorphous drugs with an accuracy compatible with the needs of the pharmaceutical industry (albeit reliable applications of ML to specific classes of, e.g., ASDs, can be found in the recent literature[22,23,29]). However, our classification model specifically has shown a remarkable potential in identifying Class I molecules: this alone is a rather crucial achievement, in that it allows us to discard potential candidates for amorphous formulations very rapidly—without the need of involving any experimental measure at all. The—important—distinction between Class II and Class III drugs is—unsurprisingly—more difficult to pinpoint exactly, and we can only hope that this work will serve to accelerate the rate at which data re: the physical stability of amorphous drugs are being collected at present. These aspects are analyzed in more details in Sec. IV.

II. METHODOLOGY

A. The datasets

The results presented in this work have been obtained with reference to two datasets: the "Amo-Reg" dataset and the "Amo-Class" dataset. The Amo-Reg dataset has been constructed from literature data: 47 data points have been taken from Ref. 19, and 131 data points have been taken from Ref. 28. The unique entries across these two datasets total 136. For each data point, we have the molecular structure [in the form of the Simplified Molecular Input Line Entry System (SMILES)[30]] and the T_g of its amorphous phase, measured via very similar experimental protocols in Refs. 19 and 28. This is important as the glass transition is not a thermodynamic property—given its value depends on the cooling/heating rate. As such, it is key that the T_g of the different systems have been measured in a consistent fashion. Given the nature of the T_g, this dataset lends itself to regression models. The Amo-Class dataset is a subset (131 molecules) of the Amo-Reg dataset and features as the target property the crystallization class we have discussed

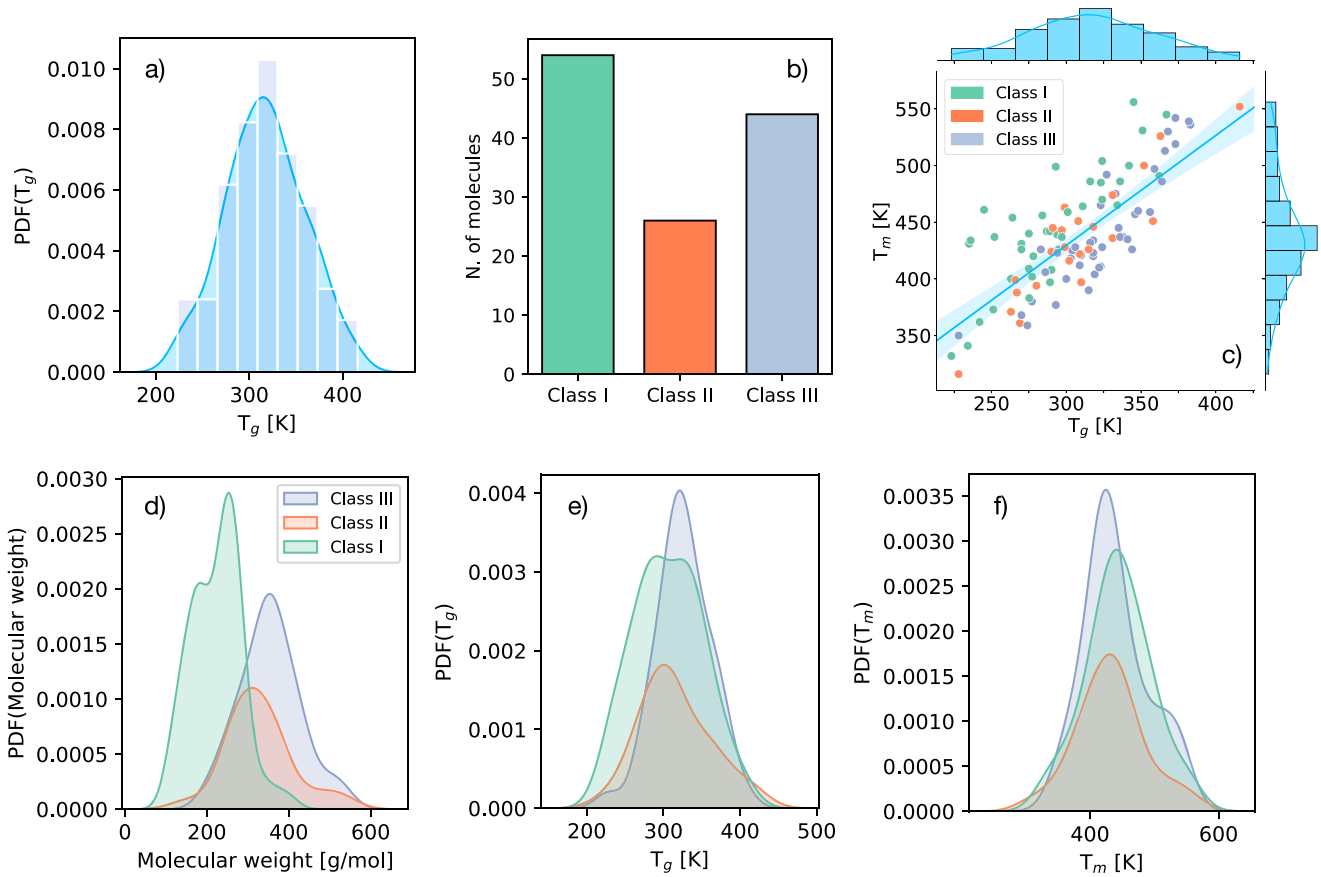

FIG. 1. The datasets. (a) Probability density function (PDF) of the T_g re: the Amo-Reg dataset (see text). The continuous distribution has been obtained via a kernel density estimation (KDE). (b) The population of each crystallization class re: the Amo-Class dataset (see text). (c) The correlation between T_g and T_m re: the Amo-Reg dataset. The marginal distributions refers to all the available data, notwithstanding the crystallization class. (d) PDFs (via KDE) of the molecular weight re: the Amo-Reg dataset. (e) PDFs (via KDE) of T_g re: the Amo-Reg dataset, with information about each different crystallization class. (f) PDFs (via KDE) of T_m re: the Amo-Reg dataset.

in the Introduction. Clearly, the nature of this dataset calls for a classification model instead.

We summarize in Fig. 1 some information about the target properties of both datasets. The distribution of T_g across the whole Amo-Reg dataset [Fig. 1(a)] appears to be peaked at around room temperature. However, the T_g of molecules belonging to Class III is rarely <250 K [Fig. 1(e)], which is consistent with the assumption that molecules characterized by high T_g are less prone to crystallize. Note that Class II is substantially under populated if compared with either Class I or III [Fig. 1(b)]—an issue we will discuss in greater detail in Sec. II C. It is also instructive to look at the distributions of the molecular weight across the different classes, reported in Fig. 1(d): specifically, Class I molecules appear to be characterized by, on average, lower molecular weight if compared to molecules in either Class II or Class III. This suggests that smaller molecules have a stronger tendency to crystallize. Intuitively, this might be explained in terms of diffusivity as smaller molecules tend to be characterized by higher self-diffusion coefficients, which, in turn, might facilitate the crystallization process (as it impacts the kinetic pre-factor re: the nucleation rate[31]). In contrast, it is challenging to extract any meaningful trend from the distributions of T_m [Fig. 1(f)], despite the fact that there exists a strong correlation between T_m and T_g, as illustrated in Fig. 1(c).

In terms of the chemical composition of the drug molecules in the dataset, we have summarized in Table I the frequency by which the relevant chemical elements appears in either the Amo-Reg or the Amo-Class datasets. The relative populations of these chemical elements are in line with those expected when considering small drug-like organic molecules.

B. Molecular dynamics simulations

One of the novel aspects of this work is the usage of MD simulations to gain access to descriptors that are simply not available when looking at single molecules in isolation. The datasets in our possession, and, indeed, the vast majority of datasets for, e.g., quantitative structure–activity relationship (QSAR) models, only provide information about the molecular structure in the form of SMILES strings. Through OpenBabel,[32] we have added hydrogen atoms where needed and generated a 3D model for each drug molecule within the Amo-Reg dataset. Note that the Amo-Class dataset is a subset of the Amo-Reg dataset (see Sec. II A). From the resulting .mol2 files, we have used CGenFF (version 4.6) to obtain the relevant topologies and force field parameters according to the CHARMM36 force field (version July 2021).[33–36]

TABLE I. Frequency by which chemical elements appears in either the Amo-Reg or the Amo-Class datasets (see text). We report the overall occurrence of a given chemical element ("Atoms" columns) as well as the number of molecules containing a given chemical element ("Molecules" columns).

	Amo-Reg		Amo-Class	
Chemical element	Atoms	Molecules	Atoms	Molecules
H	2138	136	2092	131
C	2138	136	1874	131
O	410	133	355	121
N	227	98	205	89
F	48	19	38	16
Cl	43	28	40	25
S	37	31	34	28
P	1	1	1	1

The choice of the CHARMM36 force field was simply dictated by the fact that the authors are familiar with it—and that they have previously used this particular force field with good results in the context of a variety of problems, ranging from the formation of ice at biological interfaces[37–41] to molecular glasses.[42] We have no reason to believe that the CHARMM36 force field is particularly well-suited to study molecular glasses—albeit the results discussed in Sec. III definitely support this assumption. We remark that as we are interested in extracting descriptors for the purposes of building ML models, we are not necessarily interested in the accurate parameterization of every drug molecule in the dataset. Indeed, whichever the choice of the force field, we argue that optimizing in an efficient fashion the parametrization of more than 100 molecules is unfeasible in this context. Moreover, while we do have access to the penalties associated with the parameterization of each drug molecule, we have chosen to not include these in our ML models—something that can be done via assigning uncertainties to each molecule according to their penalties. Ultimately, this choice might have had some impact in terms of the predictive power of the "solid state" descriptors discussed in Sec. III, but it did not prevent us to leverage said descriptors to improve the overall accuracy of our ML models.

The GROMACS package[43] (version 5.1.4, single precision, no GPU acceleration) has been used to perform all the MD simulations reported in this work. Periodic boundary conditions have been applied along each Cartesian direction. As we assume that both liquid and glass phases will be isotropic, we adopted cubic simulation boxes, which exact dimension was dictated by the equilibrium density of each system. Broadly speaking, the extent of the edge of said cubic boxes for the glass phases we have obtained ranges between four and six nm. The cutoff for both the van der Waals and electrostatic interactions was set to 12 Å. The van der Waals interactions were switched to zero between 10 and 12 Å. The P-LINCS algorithm[44] was used to constrain the hydrogen bonds within each molecular species. A leap-frog integrator with a time step of 2 fs was used to integrate the equations of motion. We have employed the Bussi–Donadio–Parrinello thermostat[45] and the Berendsen barostat,[46] with coupling constants of 0.5 and 4 ps, respectively, to sample either the NVT and the NPT ensemble. While we appreciate that the Berendsen barostat represents a sub-optimal choice when it comes to the accurate sampling of the isothermal–isobaric ensemble, both numerous and often drastic change in terms of simulation conditions drove us to favor the robustness of the Berendsen barostat in favor of the accuracy of more sophisticated barostats.

For each drug molecule, we have constructed an initial configuration containing 216 molecules, arranged in a simple cubic lattice in such a way to avoid any overlap between them [as illustrated in Fig. 2(d)]. We have adopted this strategy (as opposed to, e.g., random arrangements at higher densities) to enhance the mobility of the molecules within the early stages of the equilibration process. The exact simulation protocol we have employed to generate the amorphous models is summarized in Fig. 2(a). We start at very high

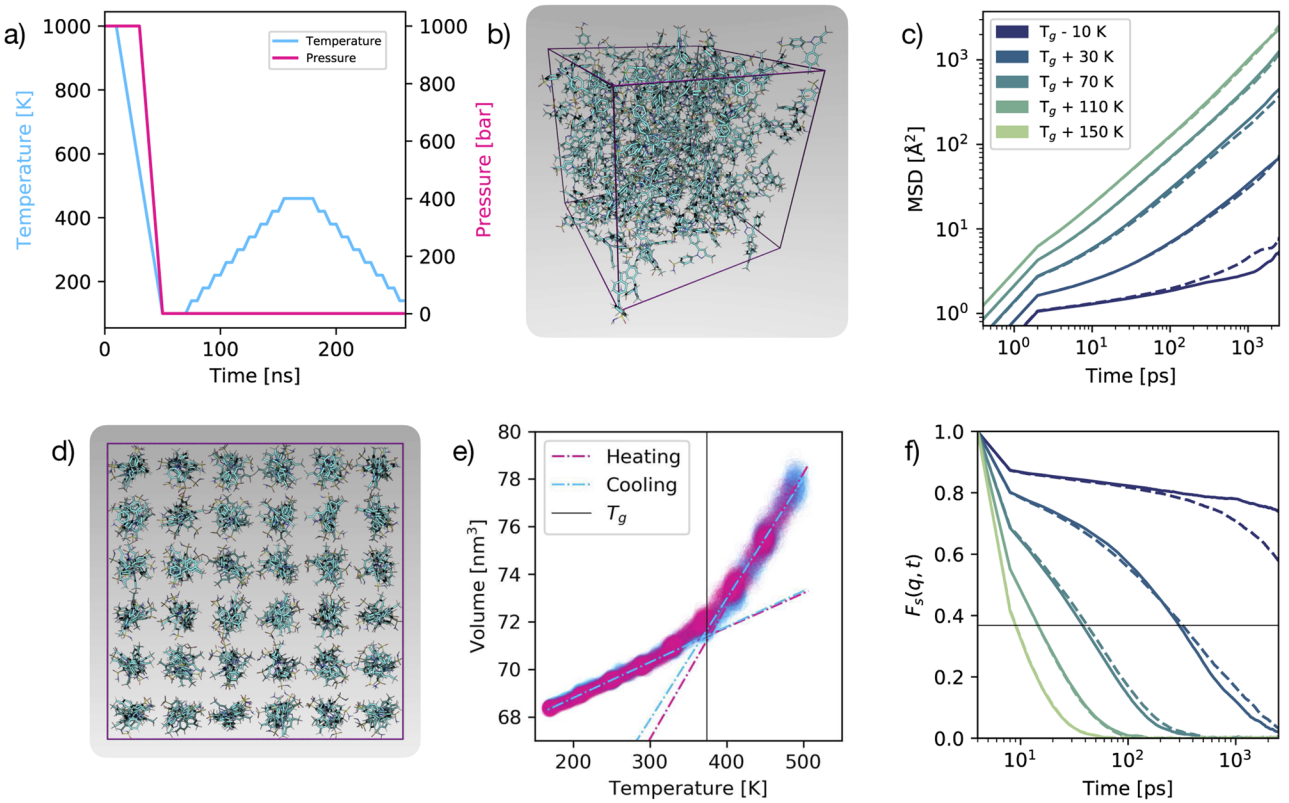

FIG. 2. Molecular dynamics simulations. (a) Computational protocol used to generate the amorphous phases of each molecule in the Amo-Reg dataset (see text). (b) A representative snapshot of an amorphous phase (for celecoxib) from a MD trajectory at 100 K. (c) Mean-squared displacement (MSD see text) for chrysin as a function of temperature—with respect to T_g. Solid and dashed lines refer to heating and cooling ramps, respectively. (d) Initial configuration for the celecoxib system. (e) Volume V as a function of temperature T for chrysin, reported along either the heating and cooling ramp. T_g can be estimated as the temperature at which the two $V(T)$ lines (obtained via linear regression, also shown) cross. (f) Self-part of the intermediate scattering function (SISF, see text) for chrysin as a function of temperature—with respect to T_g. Solid and dashed lines refer to heating and cooling ramps, respectively. The intersection between each curve and the horizontal line (at $y = 1/e$) marks the structural relaxation time τ for the system.

temperature (1000 K) and relatively high pressure (1000 bar) to randomize the system as much as possible. We then switch to ambient pressure and cool the system into a glassy state at 100 K (cooling rate: 22.5 K/ns). From there, we anneal the glass to 460 K and quench it again to 100 K (cooling rate: 4 K/ns). A representative snapshot of an amorphous phase obtained according to this protocol is shown in Fig. 2(b). Note that at this stage, we have no information about the *simulated* T_g of the system, hence why we have picked the same temperatures (100 and 460 K) for simulating the glass and the supercooled liquid phase notwithstanding the different drug molecules. Importantly, 100 and 460 K are below and above, respectively, the range of experimental T_g within our dataset.

C. Machine learning

At present, a rather diverse portfolio of ML algorithms are readily available to the scientific community. While we have experimented with several different algorithms, we have found that our results are rather robust with respect to any specific choice in that context. As such, we shall limit our discussion to neural networks (NNs) alone, which we have used in this work for both regression and classification.

In terms of the architecture of our NNs, a two-layer architecture with $2N_s + d$ nodes in each layer is, in principle, perfectly capable to deal with a dataset of N_s samples with feature dimensionality d.[47] With this in mind, we proceed with the assumption that the optimal number of nodes will be between d and $2N_s + d$. To find the number of nodes in this interval that are sufficient for our use, we multiply $2N_s + d$ by a scaling factor k, where $k \in \{0.2, 0.4, 0.6, 0.8, 1\}$. $0.2 < k < 0.8$ seems to give the best result for every descriptor we considered. We also note that descriptors characterized by high dimensionality tend to perform better with a lower scaling factor, whereas low-dimensionality descriptors prefer larger values of k. As we have found that different scaling factors do make a significant impact on the performance of several descriptors, we have used different scaling factors for each different descriptor.

The Rectified Linear Unit (ReLU) function has been utilized as the activation function within the hidden layers of our NNs. The ReLU function is slightly more computationally efficient than, e.g., a sigmoid function. Additionally, since the ReLU function is

piece-wise linear, it is inherently easier to optimize than a non-linear function and also does not suffer from the vanishing gradient problem since the value of the node activations are proportional to the gradients.[48] When it comes to the output layer, we have used a linear activation function for the regression tasks. This is standard practice as the linear activation function can output any value (it is unbounded). For the classification tasks, it is also standard practice to use the softmax function. This function is capable of converting the outputs of the neural network into j probabilities that sum up to one, where j is the number of classes.

For simplicity and to prevent us from having yet another parameter (the NN learning rate) to tune, we have opted to use the Adaptive Moment Estimation (Adam) optimizer[49] for every NN model, since the Adam optimizer is generally regarded as one of the most robust gradient decent optimization algorithms.[50] An argument could be made that stochastic gradient descent (SGD) can generalize better than the Adam optimizer.[51] However, we have explicitly verified that this is not the case here. Concerning the objective function, we have chosen the mean squared error (MSE) and the categorical cross entropy for our regression and classification models, respectively. In order to normalize the values of our features, we have used a min–max scaler.

Feature selection is often an overlooked aspect when developing machine learning models, particularly in QSAR models. It is tempting, given the availability of so many different molecular descriptors, to leverage as many of them as possible: for instance, the DRAGON software[52] can calculate more than 4800 descriptors.[53] As such, this approach is not only incredibly simple these days, but it may also enhance the flexibility of the ML algorithm of choice, in that the more the descriptors we add into the mix, the higher the chances to include those features that are actually of relevance to improve the predictive capabilities of the framework.[54] However, this strategy suffers from at least two major issues: (1) redundancy/correlation: the more the descriptors we choose to use, the higher the chance they will feed similar if not identical information to the ML algorithm,[55] with the risk of introducing artificial noise that can be detrimental to both the accuracy and the reliability of the predictive framework; (2) lack of transparency:[56,57] it becomes quite challenging to pinpoint the structural features that have the largest impact on the functional properties of interest. While from a purely practical perspective one may not care about this pitfall, understanding the structure–function relation is key to achieve the truly rational design of the novel generation of drugs.[58]

Both redundancy and lack of transparency can be mitigated by using feature selection.[19] Feature selection is the process of reducing noise in the feature space by removing a subset of features. It is important to note that, in our case, feature selection does not just reduce noise within the feature space. Since we are working with such a small dataset, we must be conscious of the dimensionality of the features we are using. For many of our descriptors, it is possible that the dimensionality of the features they return can be significantly greater than the number of data points in the training set. For this reason, the curse of dimensionality becomes a concern, and preventative measures, such as feature selection, must be taken to avoid this occurrence. To this end, we have trailed different feature selection strategies—the most effective of which turned out to be backward feature elimination. We start by using all the features available to obtain a measure of how well each individual feature can evaluate the target. Each variable is assigned a score based on how well they do at this. The lowest scoring feature is then iteratively removed until a stopping criteria is met.

As we are dealing with very small datasets, we have chosen to use the so-called leave-one-out cross validation (LOOCV) to ensure the robustness of our results. The LOOCV is a k-fold cross validation where k is equal to the number of points in the dataset. Thus, in our case, each molecule will form a test set by itself and every other molecule will be used to train the model. This is done for each molecule in the dataset, and the accuracy of the predictions can then be evaluated. The benefits of this strategy are that it allows us to train on a larger training set, mitigating some of the pitfalls of using a small amount of data to train. The main issue with LOOCV is the substantial computational cost—as we have to train a model for each data point.

Aside from its limited size, the major issue with the Amo-Class dataset is that the populations of the three crystallization classes are severely imbalanced [see Fig. 1(b)], particularly with respect to Class II molecules (which are substantially under-represented compared to the other two classes). To mitigate this drawback, we adopted the Synthetic Minority Over-sampling Technique (SMOTE, see, e.g., Ref. 59), which is a method for generating synthetic data in a relatively simple way for classification datasets. This method does not learn any underlying distribution, and it does not verify that the generated data are physically viable. However, this is not an issue, in that SMOTE is used on individual descriptors instead of the underlying dataset, and as such, it generates artificial descriptors as opposed to artificial molecules. SMOTE works on a class-by-class basis: it chooses two descriptors from the same class and draws a line between those descriptors in feature space. A point along that line is then randomly selected and that point is mapped to a descriptor vector. This is a simple and straightforward approach to data generation. It is important to note that we apply SMOTE *after* we have optimized each ML model. The newly generated features are used for training together with the original features—however, they are *not* used for predictions. This is because if we used the synthetic data we generated via SMOTE for prediction purposes, we would artificially enhance the accuracy of our ML models.

Throughout our work, we need to be able to combine different descriptors. To this end, one might be tempted to simply concatenate each feature into one (usually rather long) feature before feeding that into a single NN. However, that is a sub-optimal approach as each descriptor can be optimized by using, e.g., different regularization rates and network architectures. In order to retain the ability of optimizing each descriptor, we have leveraged ensemble methods instead. Broadly speaking, ensemble methods in machine learning are defined as the process of combining multiple models to improve the accuracy of the overall predictions. The purpose of this strategy is to reduce over-fitting, improve the extent to which the model generalizes to new data, and reduce the significance of potential outliers. First, we have attempted to train a separate NN for each descriptor. The terminal hidden layer of each NN was then merged into a single hidden layer that was then trained as a regular NN to output a single prediction. However, this approach led to only a marginal improvement of our results. This is because in merging different NNs together, we lose the ability of being able to optimize the overall NN to each descriptor. This is especially true, given that the

dimensionality of our descriptors spans several orders of magnitude, from just a single scalar (the simulated T_g) to vectors featuring up to 10^3 components (SOAP descriptors). As such, we have eventually chosen what is perhaps the simplest approach, namely, that of training models for each descriptor in isolation (thus optimizing the model specifically for a given descriptor), and then combine the predictions obtained via different descriptors. In the case of regression, this is achieved by simply taking the average of the different predictions. In the case of classification, we have adopted the widely used max-voting approach, whereby a drug molecule gets assigned to a certain crystallization class according to the highest numbers of predictions (or "votes") for that class across multiple models/descriptors. Note that this approach prevents us from using combinations of an even number of descriptors as that leads to situation where equal numbers of votes are obtained for different classes.

D. Descriptors

The molecular descriptors described in this work can be divided into two classes: "one-molecule" and "solid-state" descriptors. One-molecule descriptors are computed from the structure of a single drug molecule, considered as an isolated object in vacuum. The vast majority of descriptors utilized at present for QSAR models would fall in this category. On the other hand, solid-state descriptors are computed from the outcomes of MD simulations, which enable access to many structural and dynamical properties of the actual material. Indeed, the fundamental hypothesis at the heart of this work is that it is possible to leverage MD simulations to both complement and enhance the portfolio of traditional one-molecule descriptors.

1. One-molecule descriptors

Standard descriptors [Std]: The most rudimentary descriptor we have used is an array of what we are going to label as "standard" descriptors. These are physical properties that can easily be obtained via many different software packages, such as the RDKit[60] Python package. This package gave us access to a set of 170 2D and 3D descriptors calculated from only the molecular SMILES. In order to make use of the 3D descriptors, we had to generate 3D conformers of the molecules. We deliberately used a very basic procedure to do this whereby we leveraged the ETKDG conformation generation protocol[61] followed by UFF forcefield optimization.[62] These descriptors ranged from very basic properties, such as the molecular weight and the number of hydrogen atoms within each molecule, to more complicated ones, such as the sphericity index.[63] Although a number of these parameters (such as the WHIM descriptor[64]) can be optimized, we have mimicked the minimal effort methods used in our previous work where we show that using a large number of descriptors is not as effective as using a few carefully selected descriptors.[65]

Cliques descriptors [Cliques]: A rather simplistic and yet remarkably effective descriptor we have used in the past[65] is the so-called "cliques" descriptor. This is inspired by the work of Jin *et al.*,[66] where the authors have decomposed a given molecular structure into sub-graphs ("cliques" in graph theory), thus providing a coarse-grained molecular representation. Instead of connecting these components into a tree (as it was done in Ref. 66), we have created a vocabulary of the unique cliques across the entire dataset of interest. Thus, different sets are typically characterized by cliques vocabularies of different length. Then, we index each of the cliques in the vocabulary via an integer $i = 0, 1, \ldots, N_{clq} - 1$, where N_{clq} is the total number of unique cliques in the vocabulary. Through one-hot encoding, each molecule in the dataset is converted into a vector of length N_{clq}; the value of the ith element of said vector is equal to the number of occurrences of the ith clique within that particular molecule.

In the context of natural language processing, we are thus treating the clique vocabulary as a "bag of words" to form sentences—i.e., molecules, in a similar fashion to the "bag of bonds" descriptor explored in, e.g., Ref. 67. As the meaning of a given sentence may usually be determined to a good extent from its word content alone (i.e., without considering syntax), we are assuming that the presence of the cliques alone, without any information about the order by which they appear in a given molecular structure, would be enough to allow us to establish a structure–function relation between SMILES strings and the functional property of interest. It is thus reasonable to treat the cliques as a descriptor that is looking exclusively at the "chemistry" of the molecules, in that it highlights the presence or absence of specific molecular fragments and/or functional groups as opposed to the overall structure, albeit information about the size of the molecule is indirectly contained into the cliques vector.

Histograms of atom-centered symmetry functions [H-wACSFs]: Atom-centered symmetry functions are popular three-dimensional descriptors in the context of ML-based interatomic potentials for molecular simulations (see, e.g., Refs. 68–70). While different variations of this descriptor exist, they usually comprise sets of both radial and angular symmetry functions (SFs). In a nutshell, one sits on each atom i and computes the value of (typically Gaussian) functions that depend on either $r_{ij} = |\vec{r}_j - \vec{r}_i|$ distances (radial SFs) or θ_{ijk} angles (angular SFs) between pairs or triplets of atoms—up to a certain cutoff radius R_c. The interested reader can find a thorough introduction to SFs in Ref. 71. Crucially, the original formulation of SFs[72] required a distinct set of SFs for each combination of the different chemical species within a given molecule. While this is a perfectly sensible option in most materials science applications, where the number of elements involved is usually well below five (in fact, it is incredibly challenging to build ML-based interatomic potential for multi-component systems[68,73,74]), in the context of drug design and discovery, a molecular dataset may very well contain more than ten elements, which leads to a huge number of SFs. Gastegger *et al.* recently devised[75] a clever workaround to this issue by introducing so-called *weighted* SFs where element-dependent weighting functions depending on the atomic weight of a given atom are used to eliminate the need for separate sets of SFs for each combination of different elements, thus massively reducing the number of SFs needed as a whole.

Even weighted SFs, however, suffer from an issue of consistency, in that molecules with different elements and/or number of atoms are characterized by different numbers of SFs. As a result, the SF vectors we would like to use as inputs for our ML algorithms are not of the same length. This problem may be circumvented in several ways, none of them optimal. As a start, if one seeks to predict a functional property that can be written as the sum of atomic contributions, the original approach of Behler and Parrinello[72] can

be straightforwardly used. However, while one can think of some thermodynamic quantities, such as energy or enthalpy as additive, functional properties, or biomedical activities can often not be treated as such. Here, we have decided to build histograms of weighted-SFs (H-wACSFs): by binning the values of all the weighted SFs for each molecule, we obtained a representation that is independent from the number of atoms in a given molecule. Note that we have considered every atomic species included in Table I. While the number of bins is one of the parameters we seek to optimize (see the following section), broadly speaking low and high numbers of bins provide more or less coarse-grained representations of the molecular structure. This interesting feature can be easily leveraged in the context of three-dimensional models of crystalline or amorphous drugs—where we believe that materials science-inspired descriptors, such as H-wACSFs, could deliver important contributions.

SOAP descriptors [SOAPs]: One descriptor that in many cases has proven to be self-sufficient in offering an accurate representation of any given molecular structure is the Smooth Overlap of Atomic Potential (SOAP) descriptor,[76] even though its most commonly used form only encodes up to three-body correlations.[77] The SOAP descriptor has been gaining popularity lately, given its impressive performance across a plethora of widely different classes of materials and applications, ranging from hydrogen absorption of nano-clusters[78] to the development of bespoke interatomic potentials.[79,80] The premise of the SOAP descriptor is that, similarly to ACSFs, it offers a convenient method to describe atomic environments that are invariant to any form of rotation, translation, reflection, or permutation of equivalent atomic species. The SOAP descriptor formalism[76] leverages a set of orthonormal radial and angular basis functions to expand the local neighborhood density around each atom. An individual expansion is used for each species of atom in the neighborhood.

Several parameters can and indeed should be optimized when building a SOAP vector, namely, the number of radial basis functions, the maximum degree of the spherical harmonics, the cutoff distance for the basis function, the Gaussian smearing width of atom density, the atomic species used as centers for the basis functions, and the atomic species used as neighbors for the basis function. The optimization of these parameters is no easy feat, particularly when dealing with heterogeneous datasets. It is not obvious which sets of parameters will work when working with datasets that contain diverse molecular structures or models characterized by a variety of atomic species or environments. Initially, it may seem intuitive to simply use trial-and-error or even an exhaustive grid search strategy to optimize these parameters; however, due to the large computational costs of generating SOAP descriptors, these methods are rather inefficient. A number of approaches have been proposed in the last few years to optimize the performance of the SOAP descriptor.[76,78,79] In a recent study,[80] we have leveraged genetic algorithms (GAs)[84,85] in order to optimize the above mentioned SOAP parameters for one or multiple SOAP descriptors—given a certain choice of centers and neighbors. The very same GA-based strategy has been applied in this work to optimize the parameters of both H-wACSFs and SOAPs. Note that we have considered every atomic species included in Table I with the exception of P (only found in one instance across the whole Amo-Reg dataset).

2. Solid state descriptors

The glass transition temperature [T_g]: As discussed in Sec. I, T_g is a very important property in the context of the physical stability of amorphous drugs.[24–26] In this work, T_g is both the target property for the regression problem concerning the Amo-reg dataset (see Sec. II A) and one of the descriptors we use for the classification problem of the Amo-Class dataset (see Sec. II A), where we seek to predict which crystallization class a given drug molecule belongs to. As we have access to the experimental values of T_g for all the molecules in the Amo-reg dataset, it is instructive to compare those values with the estimates of T_g that we can obtained from our MD simulations.

To do so, we look at thermal expansion and to be specific at the volume V of the system as a function of temperature T, as illustrated in Fig. 2(e). In the temperature ranges corresponding to either the glass and supercooled liquid well below or above T_g, V is a linear function of T. In fact, the intersection between the two $V(T)$ linear regressions for the glass and supercooled liquid provides an estimate of T_g. Note that we can consider the volume variations with temperature during either the cooling or the heating ramps described in Sec. II B. This gives us two estimates for T_g, which should, in principle, be identical—but due to the thermal hysteresis caused by our rather rapid cooling/heating rates gives us a measure of the uncertainty associated with our estimate and an indication as whether our simulation protocol can be considered sufficiently accurate. Importantly, once we have obtained an estimate for the simulated T_g for each drug, we can revise our heating and cooling ramps so as to sample different systems at temperatures equally distant from their T_g. This is important because, as explained in Secs. III A–III C, we need to be consistent when producing dynamical descriptors across different drugs.

The self-diffusion coefficient [D]: Through the simulation protocol detailed in Sec. II B, we have access to the dynamical properties of the system across a wide range of temperatures. Perhaps the most basic of such properties is the self-diffusion translational coefficient, D. The latter can be obtained via either the Green–Kubo relation involving an integral of the velocity auto-correlation function or, and this is the approach we have adopted here via the Einstein relation,

$$D = \frac{1}{2d} \cdot \lim_{t\to\infty} \frac{d}{dt} \frac{1}{N} \sum_{j=1}^{N} \langle |\boldsymbol{r}_j(t_0 + \Delta t) - \boldsymbol{r}_j(t_0)| \rangle_{t_0}, \tag{1}$$

where d is the dimensionality of the system (three in our case), N is the number of molecules in the system, $\boldsymbol{r}_j$ is the position vector of the center of mass of a given ith drug molecule, and $\langle \dots \rangle_{t_0}$ refers to the time average over different, statistically independent time origins across the MD trajectory. In our case, care must be taken to identify the linear regime of the MSD: while well above T_g, the supercooled liquid is practically in a well-behaved hydrodynamic regime where the MSD grows linearly with time across the whole timescale of the simulation (exception made for the ballistic regime at very short t, which we always ignore), the MSD is basically constant for a glass as translational motion is almost absent on our timescales. A prototypical example of the variation of the MSD as a function of temperature, with respect to the T_g of the system, is reported in Fig. 2(c).

The situation becomes problematic in the proximity of T_g, where substantial deviation from linearity can be observed as the dynamics of the system progressively slows down even within the isothermal–isobaric ensemble as we are effectively attempting to extract an equilibrium dynamical property for a system that is not truly ergodic in that regime. As we need to compute D for different temperatures and more than 100 systems, however, we are forced to apply a common criterion when it comes to apply Eq. (1): specifically, we have elected to ignore the first 500 ps and last 500 ps of the MSD when calculating the slope of the linear regime. While this choice is not perfect, the results presented in Secs. III A–III C support its validity.

The structural relaxation time [τ]: The structural relaxation time τ is a measure of the timescale required for the structure of the system to evolve to a "significant" extent. In the context of the dynamics of strongly supercooled liquids and glasses, we can identify three separate relaxation timescales: (1) the above-mentioned ballistic regime, observed for very short timescales within which molecules do not have time to even collide with each other; (2) the cage-motion/rattling regime, which is absent in the hydrodynamic regime (it is a hallmark of supercooled liquids and to an even greater extent of glassy systems) and corresponds to the timescale involved with the "rattling" of the molecules within the "cages" formed by their neighbors, and the relevant timescale is often indicated as the β-relaxation of the system; (3) the α-relaxation regime, which corresponds to a significant structural change of the system due to the molecules leaving their "cages." These three timescales can all be probed by means of the (self-part of the) incoherent intermediate scattering function, which is defined as

$$F_s(\boldsymbol{q},t) = \frac{1}{N}\left\langle \sum_{j=1}^{N} \exp[i\boldsymbol{q}(\boldsymbol{r}_j(t) - \boldsymbol{r}_j(t=t_0))] \right\rangle, \tag{2}$$

where the sum runs over the jth molecule with center of mass position $\boldsymbol{r}_j(t)$ at time t, $\langle \ldots \rangle$ denotes a time average, and $\boldsymbol{q}$ is a vector in reciprocal space. Note that in an isotropic system, $F_s(\boldsymbol{q},t)$ depends only on the magnitude q of the vector $\boldsymbol{q}$. We have explicitly verified that this is the case by computing Eq. (2) by choosing $\boldsymbol{q}$ vectors along each Cartesian direction. Also note that the intermediate scattering function we obtained from our MD simulations can, in principle, be directly compared to experimental results from, e.g., inelastic neutron or x-ray scattering measurements.

As illustrated in Fig. 2(f), for a liquid in its hydrodynamic regime [see, e.g., the data re: T = T_g + 150 K in Fig. 2(f)], there is no β-relaxation regime, and $F_s(q,t)$ decays smoothly from one to zero via a single exponential decay. For a supercooled liquid, however, the emergence of the cage-rattling motion results in a characteristic plateau of the $F_s(q,t)$ [see, e.g., the data re: T = T_g + 30 K in Fig. 2(f)], which only decays to zero from the onset of the α-relaxation regime. To extract a single metric that reflects the timescale associated with the onset of the α-relaxation regime, a common choice we have also adopted in this work is that of choosing as the structural relaxation time, τ, the time for which $F_s(q,t)$ is equal to $1/e$ = 0.368. Longer relaxation times are indicative of a slower dynamics, which, in turn, should be characteristic of a lower propensity for the drug molecules in either the supercooled liquid or the glass to crystallize. τ must be related to D in some fashion, albeit τ is a much more robust quantity for a supercooled/glassy system, where not just one D exists. In fact, one can define a D characteristic to each of the timescales we have discussed. In this work, we have attempted to focus on the D associated with the α-relaxation regime, which should be directly correlated with τ. Note that below and/or in the proximity of T_g, τ might be longer than the extent of our MD simulations [see, e.g., the data re: T = T_g −10 K in Fig. 2(f)].

Other solid-state descriptors: In this work, we focused on three specific dynamical quantities. However, it is important to note that many more solid-state descriptors are directly available as the result of MD simulations. Examples involving the structural properties of the system would include the radial distribution function, the Voronoi network (and its dual tessellation, which captures the network of empty spaces within the system), and structural descriptors, such as SOAPs, that can be modified to take into account intermolecular interactions as well. The efficiency of such descriptors will be the subject of future work.

III. RESULTS

A. One-molecule descriptors

As a first attempt to predict either the T_g or the crystallization class of the drug molecules in the Amo-Reg and Amo-Class datasets, respectively, we have used different one-molecule descriptors (see Sec. II D 1).

In Table II, we report the results we have obtained re: the prediction of T_g (Amo-Reg dataset). In particular, we report the MSE for both the training and the test sets (MSE-Train and MSE-Test, respectively). Note that due to the nature of the LOOCV (see Sec. II C), we have constructed the MSE for the test set from the difference between the experimental and predicted T_g relative to individual predictions of T_g for one molecule /model. Hence, the very large values of MSE-Test, which should be interpreted as a relative measure of the variance of our results for different descriptors as opposed to the actual uncertainty associated with our results. We also report the Pearson correlation coefficient (PCC) for both the training and the test sets (PCC-Train and PCC-Test, respectively). The lack of an error bar re: PCC-Test is due to the above mentioned nature of the LOOCV.

The Std descriptor is the most accurate overall, which is to be expected as it is, in fact, a collection of 170 different descriptors. Surprisingly, the Cliques descriptor alone, despite its simplicity, shows similar accuracy. This is important as the Cliques descriptor—in stark contrast re: e.g., the Std descriptor—is a very transparent feature that can be straightforwardly leveraged to, e.g., identify specific functional groups that have an impact on, in this case, the value of T_g. The usage of genetic algorithms (GAs) is effective in improving the accuracy of the SOAP descriptor, which is consistent with our recent results[83] but only marginally effective in doing the same for H-wACSFs. Feature selection (FS) further serves to improve the accuracy of some—but not all—descriptors. Overall, Std (with FS), Cliques, and SOAPs (with FS) provide similarly accurate predictions. We will discuss how exactly our results compare to the state-of-the-art (particularly Ref. 19) in Sec. III C.

The results we have obtained re: the prediction of the crystallization class for the drug molecules in the Amo-Class dataset are reported in Table III. In this case, we have chosen Matthew's correlation coefficient (MCC[86]) to quantify the accuracy of our ML models. The MCC is much more robust than, e.g., the traditional

TABLE II. ML predictions re: the Amo-Reg dataset: one-molecule descriptors. Accuracy of our ML models in predicting the T_g re: the Amo-Reg dataset via different one-molecule descriptors (see text).

Descriptor	MSE-Train	MSE-Test	PCC-Train	PCC-Test
Std	1455.899 ± 333.474	1625.971 ± 2 637.659	0.442 ± 0.096	0.459
Cliques	1967.728 ± 608.089	1995.273 ± 2 690.114	0.266 ± 0.212	0.393
H-wACSFs	1916.213 ± 508.731	2594.003 ± 6 422.134	0.326 ± 0.121	0.157
H-wACSFs-GA	2203.201 ± 359.928	2356.687 ± 4 482.981	0.19 ± 0.105	0.166
SOAPs	2101.346 ± 796.46	3001.488 ± 13 712.585	0.353 ± 0.079	0.285
SOAPs-GA	2063.968 ± 321.171	2237.609 ± 3 135.72	0.293 ± 0.061	0.316
		Feature selection		
Std	1331.376 ± 244.329	1458.926 ± 2095.038	0.495 ± 0.066	0.511
H-wACSFs	1818.283 ± 428.084	2446.945 ± 4816.482	0.351 ± 0.109	0.186
SOAPs	1706.709 ± 516.041	2105.706 ± 6532.711	0.407 ± 0.109	0.359

definition of classification accuracy, namely, the ratio between the number of correctly classified samples and the overall number of samples, and it is especially well-suited to deal with unbalanced datasets, such as the Amo-Class dataset.[87] Again, the absence of error bars re: the MCC for the test set is due to the nature of the LOOCV we have used. Clearly, the usage of the SMOTE oversampling technique (see Sec. II C) substantially improves the predicting capabilities of every descriptor we have considered, nearing perfect accuracy re: the training set and leading to very robust results for the test sets as well. Interestingly, the performance of each individual descriptor appears to be more uniform across the Amo-Class dataset then what we have observed for the Amo-Reg dataset (where specific descriptors performed significantly better or worse). Our best result for the Amo-Class dataset has been obtained by applying SMOTE, GA, and FS on H-wACSFs—but the resulting accuracy is not drastically superior to that of any other descriptor we have considered. This is encouraging as the prediction of the crystallization class is the most useful aspect for practical considerations related to the physical stability of amorphous drugs. Again, we will discuss how exactly our results compare to the state-of-the-art (particularly Ref. 28) in Sec. III C.

B. Solid state descriptors

Before we analyze the accuracy of the ML models we have built by using solid state descriptors, it is instructive to discuss the dynamical properties we have obtained across our datasets. It is worth pointing out that, to our knowledge, this is the first, consistent collection of dynamical properties computed via MD simulations for 100+ molecular glasses, which required almost 30 μs of simulation time (roughly 3×10^6 CPU hours).

TABLE III. ML predictions re: the Amo-Class dataset: one-molecule descriptors. Accuracy of our ML models in predicting the crystallization class re: the Amo-Class dataset via different one-molecule descriptors (see text).

Descriptor	MCC-Train	MCC-Test	MCC-Train SMOTE	MCC-Test SMOTE
Std	0.837 ± 0.101	0.573	0.999 ± 0.003	0.746
Cliques	0.982 ± 0.016	0.282	0.990 ± 0.011	0.655
H-wACSFs	0.957 ± 0.06	0.405	0.421 ± 0.023	0.579
H-wACSFs-GA	0.423 ± 0.133	0.587	0.401 ± 0.111	0.601
SOAPs	0.543 ± 0.055	0.432	0.955 ± 0.057	0.659
SOAPs-GA	0.349 ± 0.104	0.59	0.979 ± 0.048	0.694
		Feature selection		
Std	0.993 ± 0.014	0.544	0.999 ± 0.040	0.747
H-wACSFs	0.687 ± 0.133	0.49	1.000 ± 0.020	0.757
H-wACSFs-GA	0.738 ± 0.175	0.464	1.000 ± 0.075	0.746
SOAPs	0.693 ± 0.154	0.399	0.975 ± 0.057	0.694
SOAPs-GA	0.918 ± 0.093	0.418	0.964 ± 0.054	0.655

We begin by comparing the estimates of T_g we have obtained from our MD simulations (see Sec. II B) with the experimental data. The results are summarized in Fig. 3(a). The raw data [gray circles in Fig. 3(a)] are very strongly correlated (Pearson Correlation Coefficient, PCC = 0.96) with the experimental data, albeit it systematically overestimates the latter. This is to be expected as the cooling rates we are forced to use in MD simulations are orders or magnitude faster than the rates achievable experimentally. However, this discrepancy appears to be entirely systematic in nature: in fact, upon shifting the raw simulation data by −70 K, we recover a very good *quantitative* agreement with the experimental data, as illustrated in Fig. 3(a) (colored points). This result demonstrates the robustness of our MD protocol and suggests that MD can indeed be used to estimate the T_g of molecular glasses with good accuracy—once the systematic error due to the non-physically fast cooling rates characteristic of typical MD timescales has been corrected for. To be specific, the cooling rate we have used here, 4 K/ns, is several orders of magnitude higher than the cooling rates achievable by means of, e.g., conventional differential scanning calorimetry (DSC), which range between 1×10^{-8} and 1×10^{-11} K/ns.

Next, we move onto the self-diffusion coefficient D. As we now have an estimate of the T_g from our MD simulations, we can measure D for the different drugs utilizing their different T_g as our reference. For instance, with "T_g + 30 K," we label the result obtained for a MD simulations at a temperature 30 K above the T_g of each specific drug. The distribution of D across the whole Amo-Reg dataset is reported in Fig. 3(d) at different temperatures. Note that we have taken the logarithm (base 10) of the actual values for visualization purposes. As expected, the higher the temperature, the higher the diffusion coefficient, which spans almost three orders of magnitude. The opposite trend can be observed for the structural relaxation time τ, also reported as a function of temperature in Fig. 3(e). In fact, D decays exponentially as τ increases, as shown by the log–log plot of D as a function of τ reported in Fig. 3(f). While

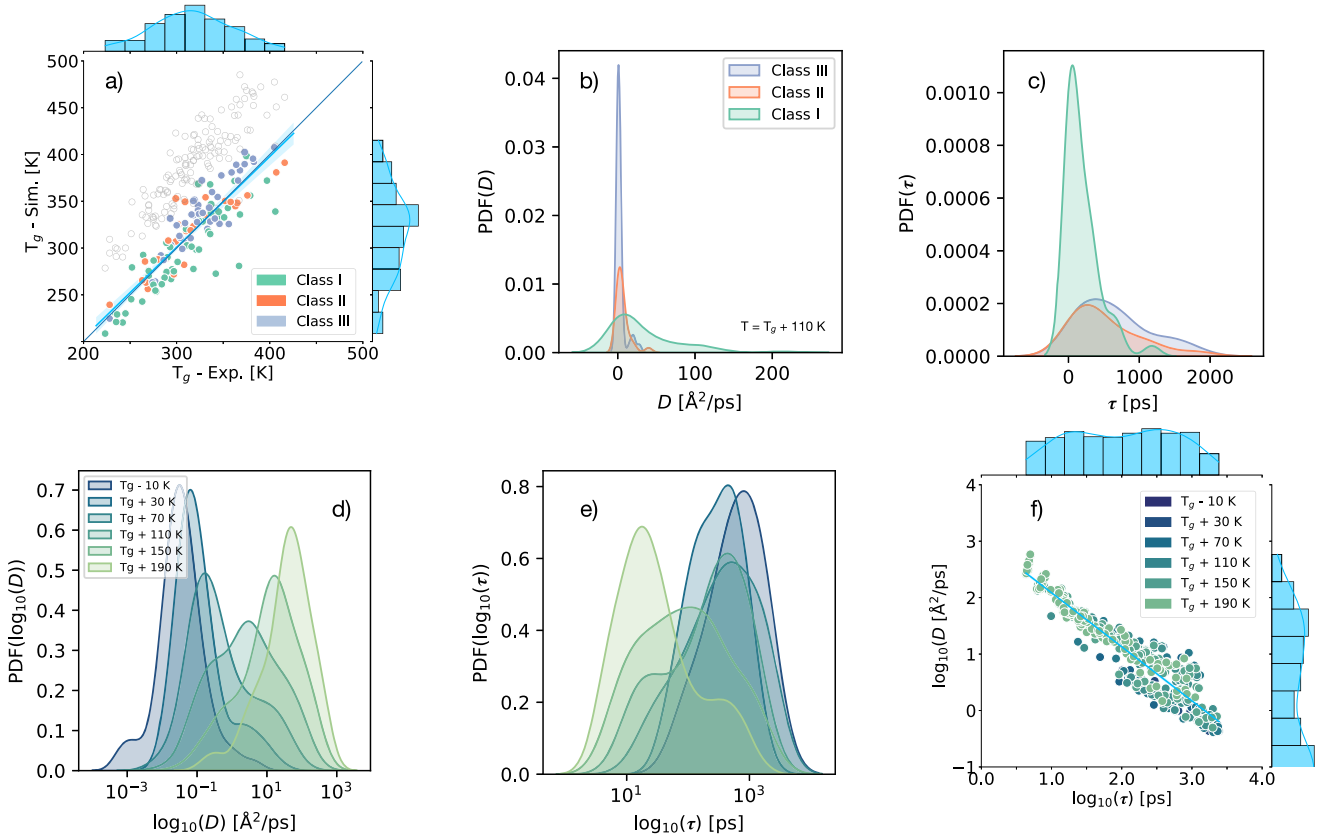

FIG. 3. Solid state descriptors. (a) Correlation between the experimental values of T_g-Exp. and the estimate of T_g obtained via our MD simulations (T_g-Sim). The raw MD results correspond to the empty, gray circles. The points colored according to the crystallization class of the corresponding molecule have been obtained by shifting the raw MD results by −70 K (see text). (b) PDF of the self-diffusion coefficient D of the supercooled liquids, labeled by crystallization class, at T = T_g + 110 K for each molecular specie in the Amo-Reg dataset. (c) PDF of the structural relaxation time τ (see text) of the supercooled liquids, labeled by crystallization class, at T = T_g + 110 K for each molecular specie in the Amo-Reg dataset. (d) PDF of the self-diffusion coefficient for each molecular specie in the Amo-Reg dataset at different temperatures. For visualization purposes, we have taken the logarithm (base 10) of the values of D. (e) PDF of the structural relaxation time for each molecular specie in the Amo-Reg dataset at different temperatures. For visualization purposes, we have taken the logarithm (base 10) of the values of τ. (f) Correlation between the (logarithm, base 10, of the) diffusion coefficient and the (logarithm, base 10, of the) structural relaxation time, color-coded according to different temperatures.

intuitive, we are not necessarily aware of a universal relation between D and τ for molecular glasses—an interesting finding in itself.

The crucial question is whether any of these dynamical quantities is related to either T_g and/or the crystallization class. Interestingly, we observe no simple relationship between either D or τ and T_g. However, there is a strong correlation between D or τ and the crystallization class, as illustrated in Figs. 3(b) and 3(c), respectively. In particular, the diffusion coefficient of class I molecules is on average much higher than class II or class III molecules. The distributions reported in Fig. 3(b) refer to the data obtained at a specific temperature above T_g (T = T_g + 110 K), but for the diffusion coefficient, this trend is present to some extent for each temperature we have considered. Note that there is no point in considering temperatures below T = T_g − 10 K as the vast majority of the glasses has impossibly low(long) diffusion coefficients(structural relaxation times). In fact, one can easily discard any drug with a $D > 50$ Å^2/ps at T = T_g + 110 K for the practical purposes of determining whether it might be suitable as an amorphous formulation. The same can be said for τ, albeit we note that the correlation between τ and the crystallization class is less strong—in that it holds to different extents according to the specific temperature chosen. These findings are of great practical relevance as they can offer a rather inexpensive route to probe whether novel candidates for amorphous drugs would fall in Class I.

Regression [T_g]: We have used these three solid state descriptors (T_g, D, and τ) to predict the experimental T_g. The results are summarized in Table IV. Unsurprisingly, given the strong correlation between simulated and experimental T_g that we have discussed above, the simulated T_g gives very accurate results: the MSE for both the training and test set is an order of magnitude lower than that obtained for any other descriptor we have used (one-molecule descriptors included, see Sec. III A).

For D and τ, we have not just one value but multiple values for each drug molecule as we have computed these dynamical quantities at several temperatures below and above their corresponding T_g. In fact, the descriptor vector for both D or τ relative to each molecule contains five elements for T_g − 10 K, T_g + 30 K, T_g + 70 K, T_g + 110 K, and T_g + 150 K. This allow us to use information about the dynamical behavior of each system as a function of temperature. While we do have data at lower/higher temperatures as well (down/up to T_g − 50 K, T_g + 190 K), virtually every system behaves in exactly the same fashion—the dynamics is either impossibly slow at T_g − 50 K or similarly fast at T_g + 190 K. As such, including D or τ for those temperatures does not increase the accuracy of our models—if anything, it introduces noise.

The results for the diffusion coefficient are as accurate as our best results we have obtained for the one-molecule descriptors upon both optimization (GAs) and feature selection—which is quite impressive. Interestingly, τ performs very poorly in terms of predicting the experimental T_g. This is not entirely unexpected as D varies much more smoothly than τ as a function of temperature. For low/high temperatures, τ tends to be either beyond the timescale that we have simulated or equally short notwithstanding the molecular species. As such, the fact that we do not observe a strong correlation between Tg and τ is probably due to the inability of our MD simulations (due to their limited timescale) to probe the full extent of τ in the proximity of T_g.

Classification [crystallization class]: We now move onto the results we have obtained with the three solid state descriptors (T_g, D, and τ) in terms of the prediction of the crystallization class of the drug molecules in the Amo-Class dataset. The results are summarized in Table V and Fig. 4. Similar to what we have observed in the case of the one-molecule descriptors, the usage of the SMOTE technique consistently improves the accuracy of our results. In terms of numerical accuracy, the results we have obtained via our solid state descriptors perform slightly worse than the one-molecule descriptors. To be specific, our best result for the one-molecule descriptors (H-wACSFs, upon FS and SMOTE, see Table III) is a MCC—for the test set—of 0.757, while our best result for the solid state descriptors (τ, upon SMOTE, see Table V) is 0.605.

In stark contrast to one-molecule descriptors, however, the solid state descriptors are characterized by a very low dimensionality—just one for Tg and five for both D and τ, while descriptors such as SOAPs can easily count 10^3 elements per SOAP vector. As such, the fact that the solid state descriptors can achieve a respectable accuracy in terms of our ML models is quite encouraging indeed. In addition to this, the solid state descriptors are very much transparent—in that they are representative of well-defined dynamical properties of the system. Again, this is a substantial advantage with respect to most one-molecule descriptors: in our case, the Cliques descriptor alone can be considered as a transparent descriptor.

Thus, it is instructive in the case of solid state descriptors to inspect the confusion matrices relative to our predictions. These are reported in Fig. 4 and refer to the results we have obtained upon applying SMOTE—for the test set only. We report a single confusion matrix per descriptor as these have been obtained via LOOCV (see Sec. II C).

We begin with T_g, which appears to be able to classify correctly the vast majority of Class II and Class III molecules, while it almost entirely mislabels Class I molecules as Class II molecules. By looking at the PDFs of the T_g for each crystallization class [see Fig. 1(e); the PDFs for the *simulated* T_g are very similar, given the strong correlation between the two, see Fig. 3(a)], the T_g of Class II and Class III

TABLE IV. ML predictions re: the Amo-Reg dataset: solid state descriptors. Accuracy of our ML models in predicting the T_g re: the Amo-Reg dataset via different solid state descriptors (see text).

Descriptor	MSE-Train	MSE-Test	PCC-Train	PCC-Test
T_g	421.759 ± 103.799	432.236 ± 902.112	0.859 ± 0.220	0.866
D	1404.177 ± 502.664	1444.206 ± 1886.394	0.487 ± 0.106	0.441
τ	1690.300 ± 233.864	1864.220 ± 2522.985	0.140 ± 0.151	0.058

TABLE V. ML predictions re: the Amo-Class dataset: solid state descriptors. Accuracy of our ML models in predicting the crystallization class re: the Amo-Class dataset via different solid state descriptors (see text).

Descriptor	MCC-Train	MCC-Test	MCC-Train SMOTE	MCC-Test SMOTE
T_g	0.241 ± 0.059	0.276	0.079 ± 0.063	0.578
D	0.271 ± 0.155	0.396	0.374 ± 0.035	0.543
τ	0.837 ± 0.101	0.573	0.207 ± 0.036	0.605

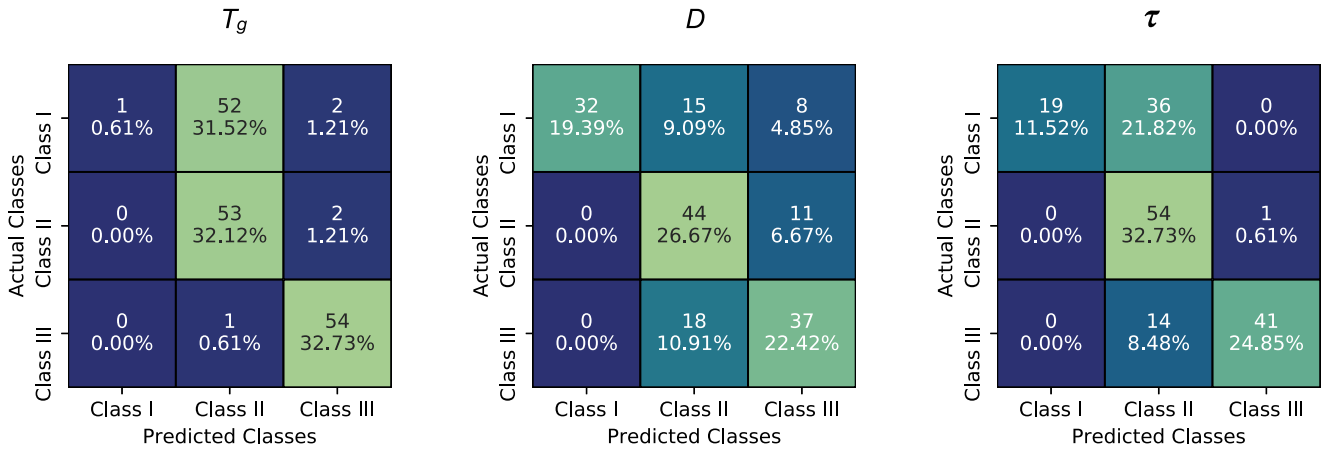

FIG. 4. Confusion matrices for the solid-state descriptors. Confusion matrices relative to our prediction of the crystallization class re: the Amo-Class dataset via different solid state descriptors. These results have been obtained upon applying the SMOTE technique and refer to the test sets only—as we adopted a LOOCV (see text).

molecules is very rarely lower than ~ 250 K. This helps in explaining why Class II and Class III are never predicted as Class I molecules in our T_g ML model, but it does not explain the close-to-perfect distinction between Class II and Class III, nor the mislabeling of almost every Class I molecule as Class II molecules.

The accuracy we have obtained when using the diffusion coefficient D as our descriptor is very similar (in terms of MCC) to that of T_g. However, as illustrated in Fig. 4, the model sacrifices some accuracy in telling apart Class II from Class III molecules to improve on the labeling of Class I molecules. Similar to what we have observed for T_g, however, no Class II or Class III molecules are ever predicted as Class I molecules.

The relaxation time τ gave us our best result in terms of accuracy. This was somehow expected as τ is a rather "binary" measure, in our case, of whether the system is behaving like a supercooled liquid (for which we observe a τ well within the timescale of our MD simulations) or a glass (in which case, we simply assign a blanket value of 9999 to the descriptor at that given temperature as we have no way to probe τ across the relevant timescale). On the contrary, τ performed rather poorly as a descriptor for the regression problem of T_g, particularly compared to D (see Table IV). The confusion matrix we have obtained for τ (see Fig. 4) is more balanced across the different classes. However, the two distinct features of all our classification models we have obtained via solid state descriptors persist, namely, (1) a substantial mislabeling of Class I molecules as Class II molecules and (2) no Class II or Class III molecules ever classified as Class I molecules.

As it stands, our models seem to be able to predict with incredible accuracy whether a drug molecule belongs to Class II or Class III—but not Class I. This is a practically important aspect. However, ideally one would be able to either "filter out" Class I molecules with great accuracy or to classify correctly Class III molecules—which are the most suitable candidates in terms of amorphous drugs formulations. For now, these models cannot achieve the former and only partially meet the expectations of the latter. We shall put our results into context re: the state-of-the-art in Sec. III C, where we will combine our descriptors into an ensemble learning framework.

C. Ensemble learning

At this stage, we have assessed the performance of a number of descriptors for both the regression problem targeting the experimental T_g and the classification problem posed by the crystallization classes. In order to increase the accuracy of our models, it is only natural to attempt to combine the predictive power of different descriptors. As discussed in Sec. II C, we have chosen to do so by training (and crucially, optimizing) models for each descriptor in isolation and combine the different predictions. The results of this ensemble approach are summarized in Fig. 5.

Regression [T_g]: We begin by discussing the results with respect to the prediction of the experimental T_g. As illustrated in Fig. 5(a), averaging our predictions over multiple descriptors does indeed lead to an improvement in the overall accuracy of our models. It is important to distinguish between two sets of results:

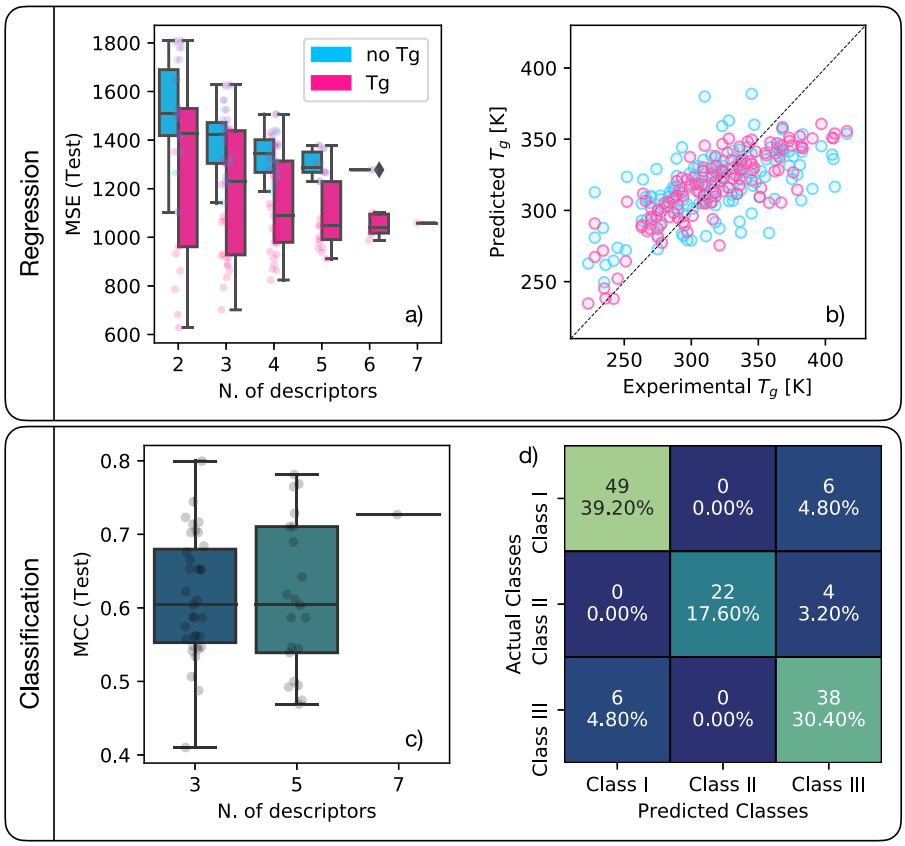

FIG. 5. Ensemble learning. (a) MSE relative to the test set for our predictions of the experimental T_g (Amo-reg dataset) as a function of the number of descriptors we have combined via ensemble learning (see text). The simulated T_g is included as a descriptor in the "Tg" results, but it has not been included in the "no Tg" results. (b) Scatter plot of the predicted vs the experimental T_g for our best regression models. The color code is the same as in panel (a). (c) MCC relative to the test set for our predictions of the crystallization class (Amo-class dataset) as a function of the number of descriptors we have combined via ensemble learning. (d) Confusion matrix for our best classification model.

ensemble models that did not include the simulated T_g as a descriptor ["no Tg" label in Fig. 5(a)] and models that did include the simulated T_g as a descriptor ["Tg" label in Fig. 5(a)]. In the latter scenario, we are effectively using a computational estimate of the target property as a descriptor—which clearly leads to much more accurate predictions.

It is interesting to look at representative scatter plots [Fig. 5(b)] of predicted vs experimental T_g for our best regression models. The "no Tg" model has been obtained by combining Std descriptors and the diffusion coefficient D. The "Tg" model has been obtained by combining the simulated T_g and the diffusion coefficient D. From these outcomes, it is evident that the diffusion coefficient brings important information about the dynamical properties of the system into the model. This is key as it demonstrates the potential of the solid state descriptors, obtained via MD simulations, which we are putting forward in this work. Interestingly, it appears that our models tend to underestimate the experimental T_g when the latter is higher than 350 K. This is somehow counter intuitive in that MD simulations systematically *over*-estimate the values of T_g [see Fig. 3(a)] due to the non-physically rapid quenching rates of the supercooled liquid into the glass.

We can now attempt to put our results into context with respect to the state-of-the-art, particularly Ref. 19. In that work, the authors considered a smaller dataset—71 molecules, to be compared with the 136 molecules in our Amo-reg dataset (see Sec. II A for further details). The authors have reported their results re: a single, specific test set of 24 molecules (which implies that they have used a single, specific training set containing 47 molecules). We believe that this approach is not ideal, in that—particularly given the small size of the dataset—the variability associated with the choice of a specific test set is bound to be very large indeed. In particular, the best result reported in Ref. 19 might have been obtained with an especially "unlucky" or even especially "lucky" test set, which prevents us from a direct comparison with our results. In contrast, by adopting the LOOCV approach we have used in this work, we are confident that our results are independent on the choice of a specific test set.

With this in mind, the MSE re: the test set obtained in Ref. 19 is 686.44. This number refers to a model leveraging NNs in a similar fashion to our work using a set of "molecular descriptors"—none of which have been obtained from the outcomes of MD simulations. As such, the descriptor we used, which is closer to the set of descriptors

used in Ref. 19, is the Std descriptor (see Sec. III A). The PCC relative to the very same model in Ref. 19 is 0.78. In comparison, our best regression model [see Fig. 5(b)] including the simulated T_g as a descriptor gave a MSE for the test set of 628.47 and a PCC of 0.83. Our best regression model [see Fig. 5(b)] obtained without including the simulated T_g as a descriptor gave a MSE for the test set of 1102.50 and a PCC of 0.60. In Ref. 19, we can also find a more accurate result (MSE = 349.69 and PCC = 0.88) obtained via support vector regression (SVR). At this stage, we do not have a robust explanation as to why SVR would perform significantly better than NNs for this particular problem. It is possible that the limited size of the dataset might be less problematic for SVR to deal with if compared to NNs, but this hypothesis will need to be verified by exploring alternative ML algorithms in the future.

Overall, our results in terms of regression are comparable to that of Ref. 19. Our combination of one-molecule and solid state descriptors gave only slightly more accurate results than the set of molecular descriptors used in Ref. 19—if we are to compare these in the context of NNs only. However, the results obtained in Ref. 19 via SVR appear to be significantly more accurate than ours, albeit it is difficult to make direct comparisons, given both the different datasets and the choice of using a single, specific test set.

Classification [crystallization class]: We have seen in Sec. III B that the relaxation time τ showed significant potential as a descriptor to identify the crystallization class. As such, it comes as no surprise that our best classification ensemble models, obtained via the straightforward max-voting approach described in Sec. II C, would include τ. The confusion matrix relative to our best classification model, which has been obtained by combining τ, wACSFs, and Std, is reported in Fig. 5(d). Note that, overall, combining multiple descriptors together did consistently improve the accuracy of our models [see Fig. 5(c)]. In contrast to the results obtained with the solid state descriptors in isolation (see Fig. 4), Class II molecules appear to be the most problematic ones to label correctly. This is expected—as we discussed in Secs. I and II A.

We can now compare our results with those of Refs. 28, where the dataset utilized (131 molecules) is identical to our Amo-Class dataset. There are two main differences between the approach of Ref. 28 and our work. The first one is that, in a similar fashion to Refs. 19, the authors have opted for a single, specific test set, which in this case encompasses 31 molecule (while the training set includes 100 molecules). Again, we do not believe this approach to be ideal. The second difference is that, in Refs. 28, the authors do not take into account Class I molecules at all for the purposes of their classification models. This has been justified on the basis of a fixed threshold re: the molecular weights of the molecules. Specifically, the authors argue that drug molecules characterized by a molecular weight (MW) <200 g/mol can be considered as Class I molecules. While it is definitely true that a strong connection between MW and crystallization class exists [see Fig. 1(d)], only 20 molecules belonging to Class I have a MW <200 g/mol. The remaining 35 Class I molecules are all characterized by MW > 200 g/mol. As such, we believe that it is not fair to exclude Class I molecules from a classification model as the MW criterion is not robust enough to provide a practical indication in terms of choosing a given drug-like molecule as a potential candidate for an amorphous formulation.

In Ref. 28, the authors have built a model (based on decision trees) that distinguish between Class II and Class III molecules (as we discussed, Class I molecules have not been taken into account) with an accuracy of 69% relative to the single, fixed test discussed above. Given than Class II is severely under-populated with respect to Class III, we argue that the usage of a metric such as the MCC we have employed here gives a better representation of the accuracy of the model. Nevertheless, for the purposes of comparing the results of our best model with that of Ref. 28, the model which confusion matrix is reported in Fig. 5(d) predicts Class I, II, and III molecules with an accuracy of 89%, 84% and 86%, respectively. Not only these numbers indicate a substantially more accurate model, but—crucially—our classification model does include Class I molecules as well.

Overall, our results in terms of classification are very encou raging—albeit it needs to be said that in order to achieve a truly predictive model, additional experimental data are certainly needed. The fact that our model identifies Class I molecules with an accuracy of almost 90% is especially intriguing for practical purposes—as we shall discuss in greater detail in Sec. IV.

IV. DISCUSSION AND CONCLUSIONS

Packaging drug molecules as amorphous solids represents an intriguing possibility for the pharmaceutical industry to circumvent the long-standing issue of the low solubility of traditional crystalline formulations. One of the major hurdles in implementing this approach, however, is the physical stability of the amorphous phase—i.e., the timescale required for it to transition into the crystal. Clearly, in order for an amorphous formulation to be marketable, the amorphous phase needs to be stable for the entire shelf life of the product—which is very difficult to predict *a priori*.

Machine learning can help in this context by developing models capable to predict the stability of a given drug molecule in its amorphous phase. However, the limited data in our possession only enable us to devise classification models aimed at predicting the so-called "crystallization class" relative to a given molecule. Classes I, II, and III correspond, loosely, to classes of drug-like molecules with very low, intermediate, and rather high physical stability. In this work, we build on the datasets and the previous results of Alhalaweh *et al.* (particularly Refs. 19 and 28) to deliver regression and particularly classification models that represent a step forward re: the state-of-the-art in terms of both accuracy and reliability. In particular, we adopt an approach that leverages the outcomes of molecular dynamics simulations to build bespoke descriptors that can be used to complement the picture offered by the traditional, "one-molecule" descriptors commonly used in QSAR models.

We combine these "solid state" descriptors with an array of optimization strategies, including genetic algorithms, feature selection, over-sampling, and ensemble learning, to craft a portfolio of classification models that—despite the very limited size of the dataset at our disposal—can correctly label drug-like molecules as Classes I, II, and III with accuracies of ~85%. The ability of our models to "filter out" Class I molecules—which are unsuitable as candidates for amorphous formulations—is especially intriguing from a practical standpoint. The outcomes of our work demonstrate the usefulness of combining molecular simulations with traditional machine learning approaches, not only to increase the predictive power of the latter but also to enable the usage of more transparent descriptors that can effectively be used to build genuine structure–function relationships between molecular structure and

functional properties. Much work remains to be done in the context of machine learning to predict the physical stability of amorphous drugs. As already mentioned, the limited size of the datasets available severely limits the accuracy of our models. We do hope that the encouraging results we have obtained here will motivate further experimental measurements aimed at increasing the size of said datasets.

It is also worth noting that many amorphous drugs are packaged as amorphous solid dispersions (ASDs): these are heterogeneous systems including amorphous drugs dispersed in (typically) polymeric matrices. Expanding the scope of our models to take into account such systems in the future is also a possibility, given that ML has been recently applied to predict the physical stability of a variety of different ASDs.[22,23,29]

In summary, we have advanced the state-of-the-art by bringing molecular simulations into the mix: at this stage, while further computational improvement is certainly possible, we believe that any decisive step forward in the field can only be achieved in conjunction with bespoke experimental efforts.

ACKNOWLEDGMENTS

T.B. thanks the EPSRC for a studentship through the Center for Doctoral Training in Mathematics for Real-World Systems II (Grant No. EP/S022244/1). T.B. and G.C.S. also gratefully acknowledge the use of the high-performance computing facilities provided by the Scientific Computing Research Technology Platform (SCRTP) at the University of Warwick.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to declare.

Author Contributions

Trent Barnard: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Gabriele C. Sosso**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data generated by the simulations reported in this manuscript are openly available at http://wrap.warwick.ac.uk.

REFERENCES

[1] E. Hadjittofis, M. A. Isbell, V. Karde, S. Varghese, C. Ghoroi, and J. Y. Y. Heng, "Influences of crystal anisotropy in pharmaceutical process development," Pharm. Res. **35**, 100 (2018).
[2] S. Datta and D. J. W. Grant, "Crystal structures of drugs: Advances in determination, prediction and engineering," Nat. Rev. Drug Discovery **3**, 42–57 (2004).
[3] S. Kalepu and V. Nekkanti, "Insoluble drug delivery strategies: Review of recent advances and business prospects," Acta Pharm. Sin. B **5**, 442–453 (2015).
[4] T. Loftsson and M. E. Brewster, "Pharmaceutical applications of cyclodextrins: Basic science and product development," J. Pharm. Pharmacol. **62**, 1607–1621 (2010).
[5] R. Laitinen, K. Löbmann, C. J. Strachan, H. Grohganz, and T. Rades, "Emerging trends in the stabilization of amorphous drugs," Int. J. Pharm. **453**, 65–79 (2013).
[6] S. B. Murdande, M. J. Pikal, R. M. Shanker, and R. H. Bogner, "Solubility advantage of amorphous pharmaceuticals, Part 3: Is maximum solubility advantage experimentally attainable and sustainable?," J. Pharm. Sci. **100**, 4349–4356 (2011).
[7] G. Chawla and A. K. Bansal, "A comparative assessment of solubility advantage from glassy and crystalline forms of a water-insoluble drug," Eur. J. Pharm. Sci. **32**, 45–57 (2007).
[8] B. C. Hancock and M. Parks, "What is the true solubility advantage for amorphous pharmaceuticals?," Pharm. Res. **17**, 397–404 (2000).
[9] C. Leuner and J. Dressman, "Improving drug solubility for oral delivery using solid dispersions," Eur. J. Pharm. Biopharm. **50**, 47–60 (2000).
[10] T. Vasconcelos, S. Marques, J. das Neves, and B. Sarmento, "Amorphous solid dispersions: Rational selection of a manufacturing process," Adv. Drug Delivery Rev. **100**, 85–101 (2016).
[11] D. Craig, P. G. Royall, V. L. Kett, and M. L. Hopton, "The relevance of the amorphous state to pharmaceutical dosage forms: Glassy drugs and freeze dried systems," Int. J. Pharm. **179**, 179–207 (1999).
[12] E. O. Kissi, H. Grohganz, K. Löbmann, M. T. Ruggiero, J. A. Zeitler, and T. Rades, "Glass-transition temperature of the β-relaxation as the major predictive parameter for recrystallization of neat amorphous drugs," J. Phys. Chem. B **122**, 2803–2808 (2018).
[13] A. D. Phan, K. Wakabayashi, M. Paluch, and V. D. Lam, "Effects of cooling rate on structural relaxation in amorphous drugs: Elastically collective nonlinear Langevin equation theory and machine learning study," RSC Adv. **9**, 40214–40221 (2019).
[14] S. Huang and R. O. Williams, "Effects of the preparation process on the properties of amorphous solid dispersions," AAPS PharmSciTech **19**, 1971–1984 (2018).
[15] X. Ma and R. O. Williams III, "Characterization of amorphous solid dispersions: An update," J. Drug Delivery Sci. Technol. **50**, 113–124 (2019).
[16] A. Singh and G. Van den Mooter, "Spray drying formulation of amorphous solid dispersions," Adv. Drug Delivery Rev. **100**, 27–50 (2016).
[17] M. Myślińska, M. W. Stocker, S. Ferguson, and A. M. Healy, "A comparison of spray-drying and co-precipitation for the generation of amorphous solid dispersions (ASDS) of hydrochlorothiazide and simvastatin," J. Pharm. Sci. (in press) (2023).
[18] J. E. Patterson, M. B. James, A. H. Forster, R. W. Lancaster, J. M. Butler, and T. Rades, "The influence of thermal and mechanical preparative techniques on the amorphous state of four poorly soluble compounds," J. Pharm. Sci. **94**, 1998–2012 (2005).
[19] A. Alzghoul, A. Alhalaweh, D. Mahlin, and C. A. S. Bergström, "Experimental and computational prediction of glass transition temperature of drugs," J. Chem. Inf. Model. **54**, 3396–3403 (2014).
[20] D. Mahlin and C. A. S. Bergström, "Early drug development predictions of glass-forming ability and physical stability of drugs," Eur. J. Pharm. Sci. **49**, 323–332 (2013).
[21] K. Nurzyńska, J. Booth, C. J. Roberts, J. McCabe, I. Dryden, and P. M. Fischer, "Long-term amorphous drug stability predictions using easily calculated, predicted, and measured parameters," Mol. Pharm. **12**, 3389–3398 (2015).
[22] R. Han, H. Xiong, Z. Ye, Y. Yang, T. Huang, Q. Jing, J. Lu, H. Pan, F. Ren, and D. Ouyang, "Predicting physical stability of solid dispersions by machine learning techniques," J. Controlled Release **311-312**, 16–25 (2019).
[23] H. Lee, J. Kim, S. Kim, J. Yoo, G. J. Choi, and Y.-S. Jeong, "Deep learning-based prediction of physical stability considering class imbalance for amorphous solid dispersions," J. Chem. **2022**, e4148443.
[24] M. Chanda, *Introduction to Polymer Science and Chemistry: A Problem-Solving Approach* (CRC Press, 2006).

[25]A. Alhalaweh, A. Alzghoul, D. Mahlin, and C. A. S. Bergström, "Physical stability of drugs after storage above and below the glass transition temperature: Relationship to glass-forming ability," Int. J. Pharm. **495**, 312–317 (2015).
[26]S. Baghel, H. Cathcart, and N. J. O'Reilly, "Polymeric amorphous solid dispersions: A review of amorphization, crystallization, stabilization, solid-state characterization, and aqueous solubilization of biopharmaceutical classification system class II drugs," J. Pharm. Sci. **105**, 2527–2544 (2016).
[27]J. A. Baird, B. Van Eerdenbrugh, and L. S. Taylor, "A classification system to assess the crystallization tendency of organic molecules from undercooled melts," J. Pharm. Sci. **99**, 3787–3806 (2010).
[28]A. Alhalaweh, A. Alzghoul, W. Kaialy, D. Mahlin, and C. A. S. Bergström, "Computational predictions of glass-forming ability and crystallization tendency of drug molecules," Mol. Pharm. **11**, 3123–3132 (2014).
[29]J. Jiang, A. Lu, X. Ma, D. Ouyang, and R. O. Williams, "The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion," Int. J. Pharm.: X **5**, 100164 (2023).
[30]D. Weininger, "Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[31]G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, "Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations," Chem. Rev. **116**, 7078–7116 (2016).
[32]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," J. Cheminf. **3**, 33 (2011).
[33]K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov *et al.*, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," J. Comput. Chem. **31**, 671–690 (2010).
[34]W. Yu, X. He, K. Vanommeslaeghe, and A. D. MacKerell, Jr, "Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations," J. Comput. Chem. **33**, 2451–2468 (2012).
[35]K. Vanommeslaeghe and A. D. MacKerell, Jr, "Automation of the CHARMM general force field (CGENFF) I: Bond perception and atom typing," J. Chem. Inf. Model. **52**, 3144–3154 (2012).
[36]K. Vanommeslaeghe, E. P. Raman, and A. D. MacKerell, Jr, "Automation of the CHARMM general force field (CGENFF) II: Assignment of bonded parameters and partial atomic charges," J. Chem. Inf. Model. **52**, 3155–3168 (2012).
[37]F. Bachtiger, T. R. Congdon, C. Stubbs, M. I. Gibson, and G. C. Sosso, "The atomistic details of the ice recrystallisation inhibition activity of PVA," Nat. Commun. **12**, 1323 (2021).
[38]C. A. Stevens, F. Bachtiger, X.-D. Kong, L. A. Abriata, G. C. Sosso, M. I. Gibson, and H.-A. Klok, "A minimalistic cyclic ice-binding peptide from phage display," Nat. Commun. **12**, 2675 (2021).
[39]C. M. Miles, P.-C. Hsu, A. M. Dixon, S. Khalid, and G. C. Sosso, "Lipid bilayers as potential ice nucleating agents," Phys. Chem. Chem. Phys. **24**, 6476–6491 (2022).
[40]M. T. Warren, I. Galpin, F. Bachtiger, M. I. Gibson, and G. C. Sosso, "Ice recrystallization inhibition by amino acids: The curious case of alpha- and beta-alanine," J. Phys. Chem. Lett. **13**, 2237–2244 (2022).
[41]G. C. Sosso, P. Sudera, A. T. Backes, T. F. Whale, J. Fröhlich-Nowoisky, M. Bonn, A. Michaelides, and E. H. G. Backus, "The role of structural order in heterogeneous ice nucleation," Chem. Sci. **13**, 5014–5026 (2022).
[42]M. González-Jiménez, T. Barnard, B. A. Russell, N. V. Tukachev, U. Javornik, L.-A. Hayes, A. J. Farrell, S. Guinane, H. M. Senn, A. J. Smith, M. Wilding, G. Mali, M. Nakano, Y. Miyazaki, P. McMillan, G. C. Sosso, and K. Wynne, "Understanding the emergence of the boson peak in molecular glasses," Nat. Commun. **14**, 215 (2023).
[43]P. Bauer, B. Hess, and E. Lindahl, Gromacs 2022.2 Manual, 2022.
[44]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "Lincs: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).
[45]G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).
[46]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684–3690 (1984).
[47]C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," Commun. ACM **64**, 107–115 (2021).
[48]X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.
[49]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).
[50]S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747 (2016).
[51]P. Zhou, J. Feng, C. Ma, C. Xiong, S. C. H. Hoi *et al.*, "Towards theoretically understanding why SGD generalizes better than ADAM in deep learning," Adv. Neural Inf. Process. Syst. **33**, 21285–21296 (2020).
[52]A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," Match **56**, 237–248 (2006).
[53]I. Olier, N. Sadawi, G. R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova, and R. D. King, "Meta-QSAR: A large-scale application of meta-learning to drug design and discovery," Mach. Learn. **107**, 285–311 (2018).
[54]A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies, "How similar are similarity searching methods? A principal component analysis of molecular descriptor space," J. Chem. Inf. Model. **49**, 108–119 (2009).
[55]M. Dehmer, F. Emmert-Streib, and S. Tripathi, "Large-scale evaluation of molecular descriptors by means of clustering," PLoS One **8**, e83956 (2013).
[56]S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," Nat. Commun. **10**, 1096 (2019).
[57]D. Castelvecchi, "Can we open the black box of AI?," Nat. News **538**, 20 (2016).
[58]J. Drews, "Drug discovery: A historical perspective," Science **287**, 1960–1964 (2000).
[59]N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res. **16**, 321–357 (2002).
[60]G. Landrum *et al.*, Rdkit: Open-source cheminformatics software, 2016.
[61]S. Riniker and G. A. Landrum, "Better informed distance geometry: Using what we know to improve conformation generation," J. Chem. Inf. Model. **55**, 2562–2574 (2015).
[62]A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," J. Am. Chem. Soc. **114**, 10024–10035 (1992).
[63]R. Todeschini and V. Consonni, "Descriptors from molecular geometry," in *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes* (John Wiley & Sons, 2003), pp. 1004–1033.
[64]P. Gramatica, "WHIM descriptors of shape," QSAR Comb. Sci. **25**, 327–332 (2006).
[65]T. Barnard, H. Hagan, S. Tseng, and G. C. Sosso, "Less may be more: An informed reflection on molecular descriptors for drug design and discovery," Mol. Syst. Des. Eng. **5**, 317–329 (2020).
[66]W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," arXiv:1802.04364 [cs, stat] (2018).
[67]C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," J. Chem. Phys. **148**, 241718 (2018).
[68]G. C. Sosso, V. L. Deringer, S. R. Elliott, and G. Csányi, "Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials," Mol. Simul. **44**, 866–880 (2018).
[69]A. Singraber, T. Morawietz, J. Behler, and C. Dellago, "Density anomaly of water at negative pressures from first principles," J. Phys.: Condens. Matter **30**, 254005 (2018).

[70] J. Li, K. Song, and J. Behler, "A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry," Phys. Chem. Chem. Phys. **21**, 9672–9682 (2019).
[71] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).
[72] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[73] F. C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V. L. Deringer, G. Csányi, and S. R. Elliott, "Modeling the phase-change memory material, $Ge_2Sb_2Te_5$, with a machine-learned interatomic potential," J. Phys. Chem. B **122**, 8998–9006 (2018).
[74] V. Quaranta, J. Behler, and M. Hellström, "Structure and dynamics of the liquid–water/zinc-oxide interface from machine learning potential simulations," J. Phys. Chem. C **123**, 1293–1304 (2019).
[75] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "WACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials," J. Chem. Phys. **148**, 241709 (2018).
[76] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).
[77] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations," Phys. Rev. Lett. **125**, 166001 (2020).
[78] M. O. Jäger, E. V. Morooka, F. Federici Canova, L. Himanen, and A. S. Foster, "Machine learning hydrogen adsorption on nanoclusters through structural descriptors," npj Comput. Mater. **4**, 1–8 (2018).
[79] J. L. Priedeman, C. W. Rosenbrock, O. K. Johnson, and E. R. Homer, "Quantifying and connecting atomic and crystallographic grain boundary structure using local environment representation and dimensionality reduction techniques," Acta Mater. **161**, 431–443 (2018).
[80] M. A. Caro, "Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials," Phys. Rev. B **100**, 024112 (2019).
[81] A. Goscinski, F. Musil, S. Pozdnyakov, J. Nigam, and M. Ceriotti, "Optimal radial basis for density-based atomic representations," J. Chem. Phys. **155**, 104106 (2021).
[82] M. O. J. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, and A. S. Foster, "Machine learning hydrogen adsorption on nanoclusters through structural descriptors," npj Comput. Mater. **4**, 37 (2018).
[83] T. Barnard, S. Tseng, J. P. Darby, A. P. Bartók, A. Broo, and G. C. Sosso, "Leveraging genetic algorithms to maximise the predictive capabilities of the SOAP descriptor," Mol. Syst. Des. Eng. **8**, 300 (2022).
[84] K. De Jong, "Genetic-algorithm-based learning," in *Machine learning* (Elsevier, 1990), pp. 611–638.
[85] J. J. Grefenstette, "Genetic algorithms and machine learning," Proceedings of the sixth Annual Conference on Computational Learning Theory, 1993, pp. 3–4.
[86] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochim. Biophys. Acta, Protein Struct. **405**, 442–451 (1975).
[87] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics **21**, 6 (2020).