Title
Accelerated Design of Disordered Materials by Computational Simulation and Machine Learning

Author
Liu, Han

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Accelerated Design of Disordered Materials by Computational Simulation and Machine Learning

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Civil Engineering

by

Han Liu

2021

ABSTRACT OF THE DISSERTATION

Accelerated Design of Disordered Materials by Computational Simulation and Machine Learning

by

Han Liu

Doctor of Philosophy in Civil Engineering

University of California, Los Angeles, 2021

Professor Mathieu Bauchy, Chair

Materials modeling is revolutionizing materials discovery paradigms through rationalizing the exploration of vast material design space. In general, materials modeling is built upon certain physics laws (e.g., computational simulations) and/or experimental data (e.g., machine learning). However, the state-of-the-art materials modeling is facing two grand challenges, i.e., (i) the high complexity of physics laws that govern materials properties, and (ii) the low informativity of experimental data. In order to address the two grand challenges of materials modeling, next-generation materials modeling aims to (i) make the physics simple to facilitate physics-driven modeling, and (ii) make the data informative to facilitate data-driven modeling.

This thesis highlights the unparallel predictive power of integrating data-driven machine learning (ML) and physics-driven computational simulations to unlock a new era for materials discovery and for next-generation materials modeling:

On the one hand, ML can assist in (i) developing empirical forcefields for accurate and computationally-efficient simulations, (ii) "separating the wheat from the chaff" in large amounts of complex simulation data to gain new insights or generate new knowledge of the underlying physics governing materials behaviors, and (iii) accelerating simulations by surrogate machine learning engines. On the other hand, simulation can generate large amounts of high-fidelity data that can be used to train machine learning models, which, in turn, can be validated by simulations. Both simulations and their integration pipeline with ML can be accelerated by leveraging automated differentiable programming engines and hardware accelerators.

Overall, I envision that the "fusion" of simulations and ML models will unlock a new era in materials modeling—wherein traditional boundaries between physics and empirical models, knowledge and data, forward and inverse predictions, or experimental and simulation data would eventually fade. I hope that the present thesis will modestly contribute to stimulating new developments in that direction.

The dissertation of Han Liu is approved.

Gaurav Sant

Jaime Marian

Scott Brandenberg

Mathieu Bauchy, Committee Chair

University of California, Los Angeles

2021

To the lovely world around me,

for your warming, guiding, and encouraging.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

A long journey would not be happier without those lovely companions:

Thanks to my advisor Professor Mathieu Bauchy from every respect!

Great thanks to my Ph.D. committee: Prof. Sant, Prof. Marian, Prof. Brandenberg, and Prof. Bauchy!

To my family, friends, colleagues, and many more!

It is my pleasure to have all of you here, and I believe this would be a cherished memory for the rest of my life!

Finally, I would like to acknowledge the journals that made publication of my work possible:

- Chapter 2 is a version of: H. Liu, L. Tang, N. Krishnan, G. Sant, M. Bauchy. *Structural Percolation Controls the Precipitation Kinetics of Colloidal Calcium–Silicate–Hydrate Gels*. **Journal of Physics D: Applied Physics** 2019, 52, 315301.
  DOI: 10.1088/1361-6463/ab217b
- Chapter 3 is a version of: H. Liu, S. Dong, L. Tang, N. Krishnan, G. Sant, M. Bauchy. *Effects of Polydispersity and Disorder on the Mechanical Properties of Hydrated Silicate Gels*. **Journal of the Mechanics and Physics of Solids** 2019, 122, 555.
  DOI: 10.1016/j.jmps.2018.10.003
- Chapter 4 is a version of: H. Liu, S. Dong, L. Tang, N. Krishnan, E. Masoero, G. Sant, M. Bauchy. *Long-Term Creep Deformations in Colloidal Calcium–Silicate–Hydrate Gels by Accelerated Aging Simulations*. **Journal of Colloid and Interface Science** 2019, 542, 339.
  DOI: 10.1016/j.jcis.2019.02.022

- Chapter 5 is a version of: H. Liu, T. Zhang, N. Krishnan, M. Smedskjaer, J. Ryan, S. Gin, M. Bauchy. *Predicting the Dissolution Kinetics of Silicate Glasses by Topology-informed Machine Learning*, ***npj Materials Degradation*** 2019, 32.

  DOI: 10.1038/s41529-019-0094-1

- Chapter 6 is a version of: H. Liu, Y. Liu, Z. Zhao, S. Schoenholz, E. Cubuk, M. Bauchy. *End-to-End Differentiability and Tensor Processing Unit Computing to Accelerate Materials' Inverse Design.* **Workshop on machine learning for engineering modeling, simulation and design @ NeurIPS** 2020.

- Chapter 7 is a version of: H. Liu, Z. Fu, Y. Li, N. Sabri, M. Bauchy. *Parameterization of Empirical Forcefields for Glassy Silica Using Machine Learning*. **MRS Communications** 2019, 9, 593.

  DOI: 10.1557/mrc.2019.47

- Chapter 8 is a version of: H. Liu, Z. Fu, Y. Li, N. Sabri, M. Bauchy. *Balance between Accuracy and Simplicity in Empirical Forcefields for Glass Modeling: Insights from Machine Learning*. **Journal of Non-Crystalline Solids** 2019, 515, 133.

  DOI: 10.1016/j.jnoncrysol.2019.04.020

- Chapter 9 is a version of: H. Liu, Y. Li, Z. Fu, K. Li, M. Bauchy. *Exploring the Landscape of Buckingham Potentials for Silica by Machine Learning: Soft vs Hard Interatomic Forcefields.* **Journal of Chemical Physics** 2020, 152, 051101.

  DOI: 10.1063/1.5136041

- Chapter 10 a version of: H. Liu, E. Bao, E. Li, E. D. Cubuk, S. S. Schoenholz, S. Xiao, Z. Zhao, C. Yang, G. Sant, M. Smedskjaer, M. Bauchy. *Finding Needles in Haystacks:*

Deciphering a Structural Signature of Glass Dynamics by Machine Learning. In preparation.

- Chapter 11 a version of: H. Liu, S. Xiao, L. Tang, E. Bao, E. Li, C. Yang, Z. Zhao, G. Sant, M. Smedskjaer, L. Guo, M. Bauchy. *Predicting the Early-Stage Creep Dynamics of Gels from Their Static Structure by Machine Learning.* **Acta Materialia** 2021, 210, 116817. DOI: 10.1016/j.actamat.2021.116817

VITA

Education

2016 – 2021	Ph.D. Candidate in Civil Engineering
	M.S. in Electrical Engineering (concurrently, 2020)
	University of California, Los Angeles
	Los Angeles, US
2010 – 2016	M.S. in Materials Science (2016)
	B.S. in Business Administration (concurrently, 2014)
	B.S. in Polymer Materials and Engineering (2013)
	Sichuan University
	Chengdu, China

Publications

[14] H. Liu, S. Xiao, L. Tang, E. Bao, E. Li, C. Yang, Z. Zhao, G. Sant, M. Smedskjaer, L. Guo, M. Bauchy. *Predicting the Early-Stage Creep Dynamics of Gels from Their Static Structure by Machine Learning.* **Acta Materialia** 2021, 210, 116817.

[13] H. Liu, Y. Liu, Z. Zhao, S. Schoenholz, E. Cubuk, M. Bauchy. *End-to-End Differentiability and Tensor Processing Unit Computing to Accelerate Materials' Inverse Design.* **Workshop on machine learning for engineering modeling, simulation and design @ NeurIPS** 2020.

[12] H. Liu, Y. Li, Z. Fu, K. Li, M. Bauchy. *Exploring the Landscape of Buckingham Potentials for Silica by Machine Learning: Soft vs Hard Interatomic Forcefields.* **Journal of Chemical Physics** 2020, 152, 051101.

[11] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy. *Machine Learning for Glass Science and Engineering: A Review.* **Journal of Non-Crystalline Solids: X** 2019, 4, 100036.

[10] H. Liu, T. Zhang, N. Krishnan, M. Smedskjaer, J. Ryan, S. Gin, M. Bauchy. *Predicting the Dissolution Kinetics of Silicate Glasses by Topology-informed Machine Learning*, ***npj Materials Degradation*** 2019, 32.

[9] H. Liu, L. Tang, N. Krishnan, G. Sant, M. Bauchy. *Structural Percolation Controls the Precipitation Kinetics of Colloidal Calcium–Silicate–Hydrate Gels*. **Journal of Physics D: Applied Physics** 2019, 52, 315301.

[8] H. Liu, Z. Fu, Y. Li, N. Sabri, M. Bauchy. *Parameterization of Empirical Forcefields for Glassy Silica Using Machine Learning*. **MRS Communications** 2019, 9, 593.

[7] H. Liu, S. Dong, L. Tang, N. Krishnan, E. Masoero, G. Sant, M. Bauchy. *Long-Term Creep Deformations in Colloidal Calcium–Silicate–Hydrate Gels by Accelerated Aging Simulations*. **Journal of Colloid and Interface Science** 2019, 542, 339.

[6] H. Liu, Z. Fu, Y. Li, N. Sabri, M. Bauchy. *Balance between Accuracy and Simplicity in Empirical Forcefields for Glass Modeling: Insights from Machine Learning*. **Journal of Non-Crystalline Solids** 2019, 515, 133.

[5] H. Liu, S. Dong, L. Tang, N. Krishnan, G. Sant, M. Bauchy. *Effects of Polydispersity and Disorder on the Mechanical Properties of Hydrated Silicate Gels*. **Journal of the Mechanics and Physics of Solids** 2019, 122, 555.

[4] H. Liu, T. Du, N. Krishnan, H. Li, M. Bauchy. *Topological Optimization of Cementitious Binders: Advances and Challenges*. **Cement and Concrete Composites** 2019, 101, 5.

[3] H. Liu, G. Huang, J. Zeng, L. Xu, X. Fu, S. Wu, J. Zheng, J. Wu. *Observing Nucleation Transition in Stretched Natural Rubber through Self-seeding*. **Journal of Physical Chemistry B** 2015, 119, 11887.

[2] H. Liu, G. Huang, L. Wei, J. Zeng, X. Fu, C. Huang, J. Wu. *Inhomogeneous Natural Network Promoting Strain-induced Crystallization: A Mesoscale Model of Natural Rubber*. **Chinese Journal of Polymer Science** 2019, 37, 1142.

[1] H. Liu, M. Zhou, Y. Zhou, S. Wang, G. Li, L. Jiang, Y. Dan. *Aging Life Prediction System of Polymer Outdoors Constructed by ANN. 1. Lifetime Prediction for Polycarbonate*. **Polymer Degradation and Stability** 2014, 105, 218.

Chapter 1. Introduction

1.1 Motivation for Modeling of Disordered Materials

1.1.1 Vast Design Space of Disordered Materials

Developing novel materials with new, improved properties and functionalities is key to address some of the Grand Challenges facing our society [1,2]. Although the process of designing a new material is always a difficult task, the design of novel disordered materials (e.g., glasses) comes with some unique challenges [3]. Taking the example of glassy materials, virtually all the elements of the periodic table can be turned into a glass if quenched fast enough [4]. Moreover, unlike crystals, noncrystalline solids are intrinsically out-of-equilibrium and, hence, can exhibit a continuous range in their stoichiometry (within the glass-forming ability domain) [5]. For both of these reasons, the compositional envelope that is accessible to glass is limitless and, clearly, only an infinitesimal fraction of these compositions have been explored thus far [4].

Figure 1-1 shows an atomistic structure of sodium silicate glass prepared by melt-quenching molecular dynamics (MD) simulation. Notably, the disordered structure exhibits a variety of complexity at different length scale, including atom type, bond length, bond angle, atom coordination number, ring size, Voronoi volume, local packing density, radial 2-body order, angular 3-body order, etc. As a key advantage of disordered materials, such high structural complexity provides a vast design space to tune materials properties [6]. Although the design space accessible to disordered materials (e.g., the infinite compositional envelope to glasses) opens endless possibilities for the discovery of new structures with unusual properties [4], efficiently exploring such a high-dimension space is notoriously challenging and traditional discovery methods based on trial-and-error Edisonian approaches are highly inefficient (see Sec. 1.1.2) [7,8].

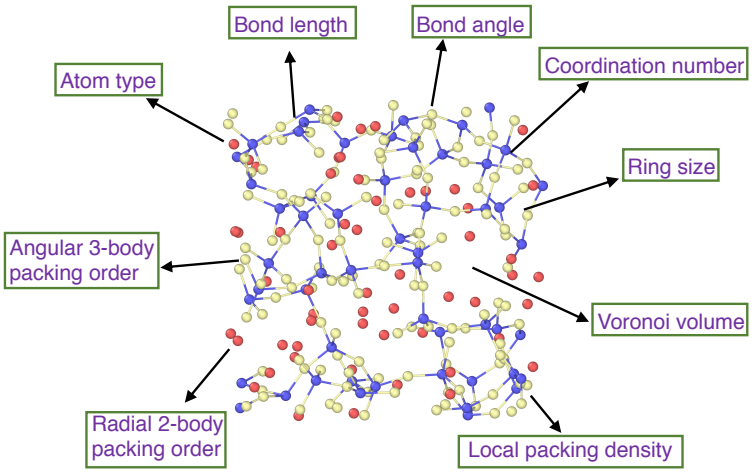


Figure 1-1: Illustration of structural complexity in a sodium silicate glass ($(Na_2O)_{30}(SiO_2)_{70}$). The glass structure exhibits various structural features, including atom type, bond length, bond angle, atom coordination number, ring size, Voronoi volume, local packing density, radial 2-body order, angular 3-body order, etc.

1.1.2 Traditional Design Paradigm by Edisonian Trial-and-Error Experiments

Due to their structural complexity (see. Fig. 1-1), disordered materials exhibit continuously changing properties in their vast design space [6]. Figure 1-2 illustrates a material design space, where the color coding represents a target property of the material. In order to find out the optimal material exhibiting optimal target property, we need to search the high-dimension design space point by point, and in our human history, numerous efforts have been made to find a more efficient search strategy for materials discovery [9,10].

The traditional paradigm for materials discovery is based on "intuition" and takes many trial-and-error "Edisonian" experiments (see Fig. 1-2) [10]. Note that, this method is widely applied in laboratory and industry [11]. Based on the accumulated data and experiences, the next candidate datapoints are usually chosen with some randomness or slightly adjusted from previous

experiments, resulting in sparse data over the entire design space or clustered data in a small region [11,12]. However, both the two types of data explorations are not ideal situation to find the optimal materials [11,12]. Moreover, traditional experiments require relatively high material and time cost to generate data points and cannot produce enough data points cover the entire exploration domain [11].

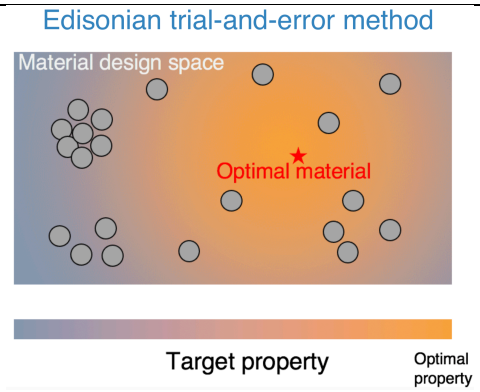


Figure 1-2: Illustration of Edisonian trial-and-error method for materials discovery. The color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by trial-and-error method.

1.1.3 Accelerated Design of Disordered Materials by *In Silico* Modeling

In order to overcome the limitations of intuition-based "trial-and-error" approach and reduce search randomness (see Sec. 1.1.2), past centuries have witnessed significant revolutions in materials discovery paradigm [9,10]. Figure 1-3 shows the evolution of materials discovery paradigms along human history timeline [10]. Along with the advancement of science and technology, the revolution for materials discovery is roughly divided into four stages, i.e., empirical, theoretical, computational, and data-driven stages [10].

In stage I (before seventeen century), materials discovery is purely driven by empirical science, and our ancestors relied on their intuitions to refine materials using Edisonian trial-and-error approach (see Sec. 1.1.2)—although it is rarely possible to find the optimal material in the vast design space [9,10]. Then starting from seventeen century (stage II), modern science emerges, and materials scientists started to propose, establish, and continuously refine a systematic set of model-based theories that govern materials properties (e.g., laws of thermodynamics) [6]. These material theories greatly facilitate materials discovery by identifying the most promising "optimal" regions in their design space [9,10]. However, there remains some level of intuition-based randomness to explore the identified local regions [13].

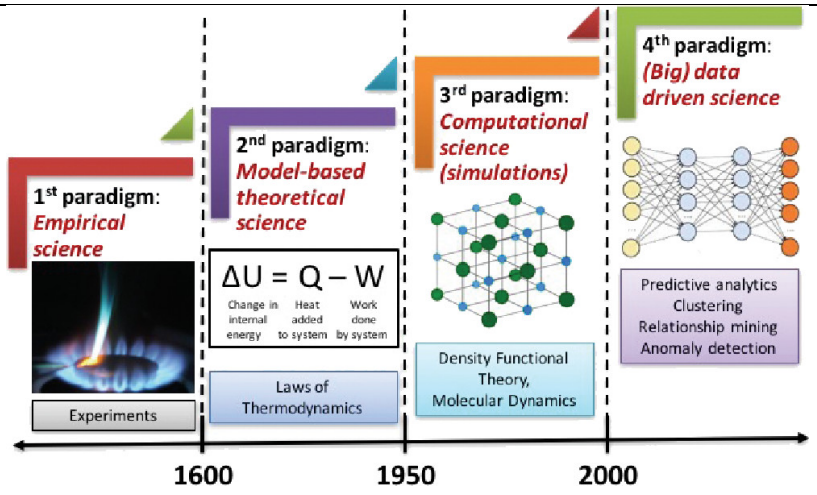


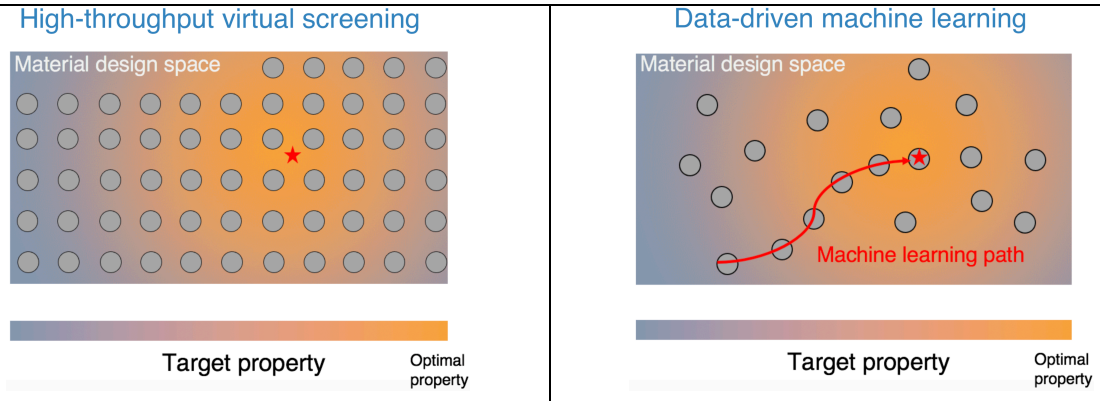
Figure 1-3: Schematic of fourth paradigm for materials discovery. Along the human history timeline, the way we discover materials is evolving from empirical (1st paradigm), to theoretical (2nd paradigm), computational (3rd paradigm), and now, to data-driven paradigm (4th paradigm). Image adopted from ref. [10]

Starting from the middle of last century (stage III), we have witnessed the revolution of every scientific field with the advancement of computer science [14]. In this stage, relying on more

and more powerful computing resources, we envisioned to encode any model-based material theories into computational algorithms predictive in materials behaviors and properties—namely, computational simulations (e.g., molecular dynamics) [15]. Compared to traditional experiments, simulations usually require relatively lower material (computational power only) and time cost, and therefore, can generate much more data points in short time [11]. As such, many materials experiments were expected to be conducted by computational simulations as a more efficient *in silico* experiments—rather than conducted by traditional laboratory experiments [16]. As an ideal scenario, high-throughput virtual screening (HTVS) has been proposed to allow computational tools to provide an efficient way to test on-the-fly every datapoint in material design space [17], as illustrated in Fig. 1-4a. Although there is no guarantee that HTVS can provide the optimal material unless more finer simulations (i.e., higher resolution over the entire design space tested by simulations), it remains a "cheap" approach to guide experiments and accelerate the discovery of new materials [17].

Now entering into this century (stage IV), the accumulation of materials data during past centuries along with the thrust of computer science has unlocked a new era of "big data" [9,10]. In this stage, materials scientists aim to leverage the power of data-driven science (e.g., machine learning) to efficiently establish data-driven (physics-blind) models that (i) predict materials properties and (ii) guide the discovery of optimal material [3]. Figure 1-4b shows an illustration of machine learning (ML) for materials discovery. In detail, using the learning examples (i.e., the training dataset), ML infers the landscape of target property over the entire material design space through a ML regression model [3]. This landscape is informative to guide a ML search path toward local regions most promising to exhibit optimal properties (see Fig. 1-4b), thus significantly reducing the exploration efforts [7,8]. Overall, in contrast to traditional laboratory "trial-and-error"

explorations, both stage III and stage IV paradigm greatly facilitate materials discovery by resorting to *in silico* modeling approaches—either driven by physics (3rd paradigm) or by data (4th paradigm) [10,18].



(a) **(b)**

Figure 1-4: Illustration of **(a)** high-throughput virtual screening and **(b)** machine learning method for materials discovery. The color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by the method, and the red arrow shows the machine learning search path.

1.2 Overview of the State-of-the-Art in Modeling of Disordered Materials

1.2.1 Materials Modeling driven by Physics and Data

Looking into the evolution of materials discovery paradigms (see Fig. 1-3) [10], it is notable that, having an extensive set of materials modeling tools revolutionizes the way we discover new materials [18]. In general, materials modeling is built upon certain physics laws and/or experimental data [18]. Figure 1-5 shows a variety of modeling tools [18], ranging from functional physics-driven models (e.g., quantum mechanics simulation [19]), to empirical models relying on both physics laws and experimental data (e.g., topological constraint theory [20]), and

to the most extreme case where the modeling is completely physics-blind and purely driven by data (e.g., data-driven machine learning [3]). In this thesis, based on whether physics law involved or not, materials modeling tools are roughly categorized into (i) physics-driven modeling (e.g., computational simulations, see Sec. 1.2.2) and (ii) data-driven/physics-blind modeling (e.g., machine learning models, see Sec. 1.2.3) [18]. Note that, the scope of this thesis would mainly focus on two representative modeling tools in the two categories, respectively, that is, (i) physics-driven computational simulations (3rd paradigm for materials discovery, see Fig. 1-4a), and (ii) data-driven machine learning (4th paradigm for materials discovery, see Fig. 1-4b).

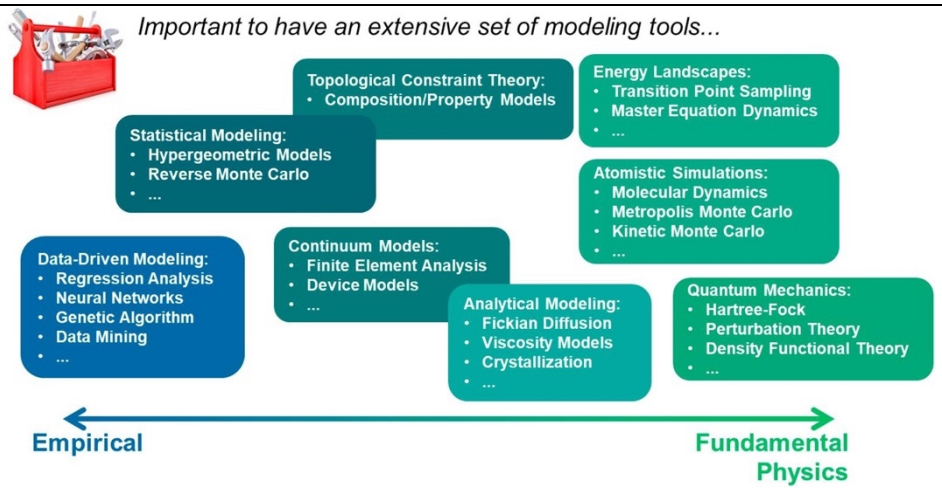


Figure 1-5: Illustration of various materials modeling tools, ranging from physics-driven modeling to (physics-blind) data-driven modeling. Image adopted from ref. [18]

1.2.2 Computational Simulation: Physics-Driven Modeling

Built on computational programming platforms (containing both software and hardware), computational simulations aim to implement model-based material theories into computational algorithms—which enable predictions of materials behaviors and properties [15]. Here, taking the example of atomistic simulations, Figure 1-6 illustrates the physics principles behind their

algorithm implementations so as to reveal the physics-driven nature of computational simulations [21]. As a simple classification, atomistic simulations can be broadly divided in terms of their description of the atomic motion [21]. Namely, (i) molecular dynamics (MD) simulations offer a direct description of the spontaneous dynamics of the atoms as per the Newton's law of motion [15], whereas (ii) other types of simulations (e.g., Monte Carlo (MC) simulations [14]) simply aim to construct an atomic structure based on a target objective (e.g., minimizing energy) without any explicit description of the dynamics of the atoms [14].

MD simulations predict the true motion of the atoms that is predicted by the Newton's law of motion (see Fig. 1-6a) [22]. This requires the knowledge of the interatomic forcefield, that is, the real-time force experienced by each atom [23]. Such forces can comprise radial 2-body interactions [24], angular 3-body interactions [25], and/or many-body interactions [26] and play a key role in predicting atom trajectories. In practice, the interatomic forcefield can be accurately computed using first-principles electron-level methods (e.g., *ab initio* MD simulation [19]) or can be approximately estimated by some empirical functions (i.e., classical MD simulation [27,28]).

Unlike MD simulations, other types of atomistic simulations such as energy-based MC [14], or energy minimization based on gradient descent [29,30] do not follow the Newton's law of motion. Rather, these approaches rely on exploring and finding the minimum position of a given "cost landscape" (e.g., potential energy for MC and energy minimization) via a series of structural modifications (e.g., displacing atoms) [22]. For instance, Figure 1-6b illustrates the principle behind energy-based MC simulations [21], wherein the MC simulation searches the global minimum within the potential energy landscape (PEL) by performing a series of tentative MC moves. Each move is either accepted or rejected according to a given acceptance probability defined in the MC algorithm [14]. In contrast to MD simulations wherein large energy barriers are

unlikely to be overcome [29,31], simulations based on sampling a PEL can "accelerate" atomic motion to (i) jump over energy barriers and (ii) move toward the minimum energy state [32,33], which corresponds to the most stable energy state that the atomic system relaxes toward upon aging/relaxation [34]. Note that the landscape to explore (i.e., the function to minimize) is not always the potential energy, but, rather, can take the form any function, e.g., a loss function capturing the structural difference between simulated and experimental glass in the case of reserve Monte Carlo (RMC) simulations [35].

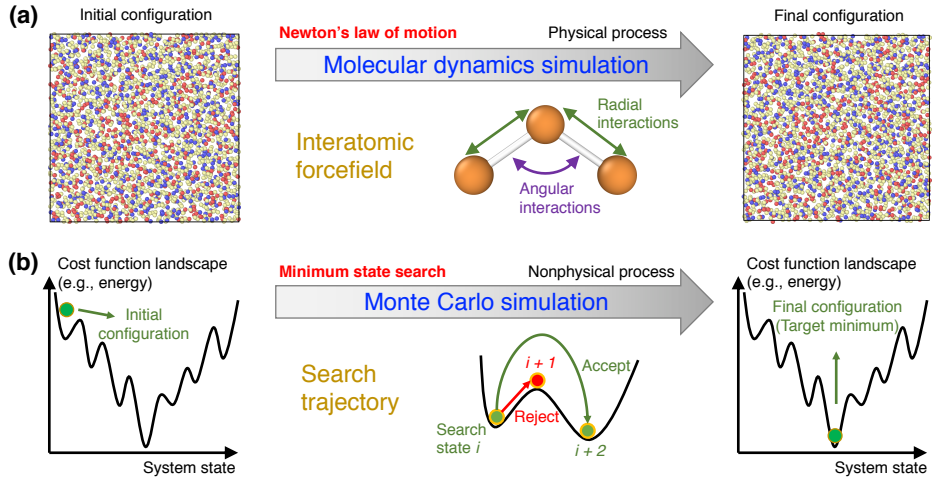


Figure 1-6: (a) Illustration of a molecular dynamics (MD) simulation of a glass system, wherein, starting from an initial configuration, the motion of the atoms is determined based on the interatomic interactions following the Newton's law of motion. **(b)** Illustration of a Monte Carlo (MC) simulation, wherein an MC search algorithm (e.g., energy-based Metropolis algorithm) is used to find the minimum state (e.g., minimum energy) of a glass system within a cost function landscape—e.g., potential energy landscape (PEL), namely, a system's potential energy as a function of its atom positions. The landscape is sampled by performing a series of MC moves (e.g., random displacement of an atom). Image adopted from ref. [21]

1.2.3 Machine Learning: Data-Driven Modeling

As an alternative route to physics-based modeling, artificial intelligence (AI) and machine learning (ML) offer a promising path to leverage existing datasets and infer data-driven models that, in turn, can be used to accelerate the discovery of novel materials [36,37]. Over the past decade, thanks to the rapid increase in available computing power [38], AI and ML have revolutionized various aspects of our lives [39], including for image recognition [40], Internet data mining [41], or self-driving cars [42].

In details, machine learning can "learn from example" by analyzing existing datasets and identifying patterns in data that are invisible to human eyes [43]. Figure 1-7 shows a typical application of machine learning to glass design [3]. First, some data are generated (by experiments, simulations, or mining from existing databases) to build a database of properties. Such databases can comprise, as an example, the glass composition, synthesis procedure, as well as select properties. Machine learning is then used to infer some patterns within the dataset and establish a predictive model.

Machine learning algorithms can accomplish two types of tasks, namely, supervised and unsupervised [44]. In the case of supervised machine learning, the dataset comprises a series of inputs (e.g., glass composition) and outputs (e.g., density, hardness, etc.). Supervised machine learning can then learn from these existing examples and infer the relationship between inputs and outputs. Supervised machine learning comprises (i) regression algorithms [3], which can be to predict the output as a function of the inputs (e.g., glass composition-property predictive models) and (ii) classification algorithms [3], which can be used to label materials within different categories (e.g., transparent or nontransparent glasses). In contrast, in the case of unsupervised machine learning, the dataset is not labeled (i.e., no output information is known). Unsupervised

machine learning can, for instance, be used to identify some clusters within existing data, that is, to identify some families of data points that share similar characteristics [3].

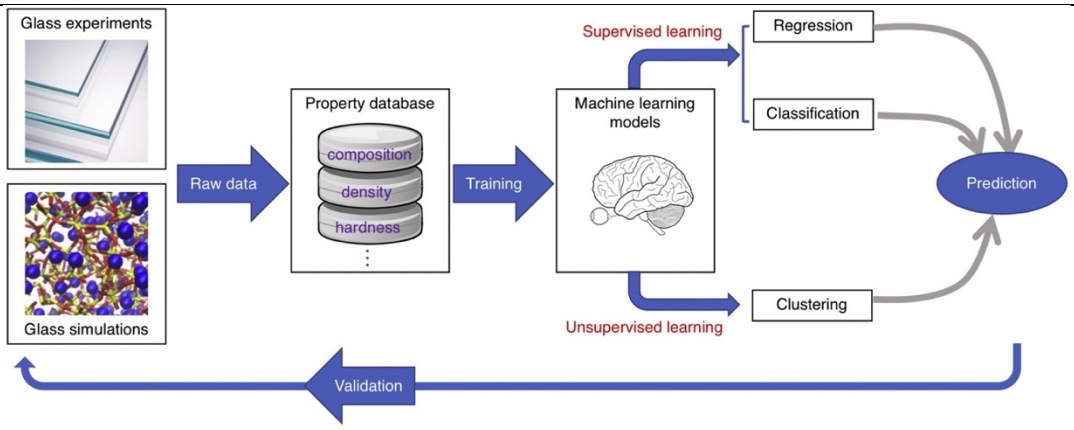


Figure 1-7: Illustration of a machine learning (ML) pipeline for glass design. ML models are generally applied to two types of learning tasks, i.e., supervised learning (e.g., regression and classification), and unsupervised learning (e.g., clustering). Image adopted from ref. [3]

1.3 Present Challenges in Modeling of Disordered Materials

1.3.1 Is the Physics Simple?

Although various materials modeling tools, including both physics- and data-driven models, have been established to accelerate materials design (see Fig. 1-5) [18,36], there remain two grand challenges facing materials modeling in general, that is, (i) the high complexity of physics laws that govern materials properties (see Fig. 1-8) [45], and (ii) the low informativity of experimental dataset (see Fig. 1-9) [11,12]. In this section, I would introduce the complex physics that challenges the development of physics-driven modeling.

In order to build a physics-based predictive model, it is necessary to quantify the physics laws governing materials properties. However, on the one hand, the underlying physics might be too complex to be simplified as quantitative model-based theories. One typical example is glass

dissolution (see Chapter 5) [12], where the dissolution rate can be divided into three stages governed by different, complex dissolution mechanisms [46]. This renders challenging to build a physics-driven model that can meaningfully predict the glass dissolution rate, and we then resort to build a data-driven model instead (see Chapter 5) [12].

On the other hand, although there exists a ground truth that nearly all condensed-state properties of materials can be ascribed to the electron-level interactions as explicitly formulated by first-principles theories [19], building a predictive model relying on first-principles formulations would result in notoriously expenesive computational cost [47,48]. Figure 1-8 illustrates a computational expensive virtual screening relying on first-principles molecular dynamics (MD) simulations [19]. MD simulation implements Newton's law of motion to predict atom motions by a loop of 4-successive-step computational algorithm [21], namely, (i) computing the system's potential energy $U(\{r_i\})$ by summing up all interatomic interactions for the current atom positions $\{r_i\}$, (ii) calculating the resultant force $\{F_i\}$ experienced by each atom i via energy differentiation (i.e., $F_i = -\partial U/\partial r_i$), (iii) obtaining each atom's acceleration $\{a_i\}$ from $\{F_i\}$ as per the Newton's law of motion, that is, $a_i = F_i/m_i$, where m_i is the mass of atom i, and finally, (iv) updating the atom positions and velocities after a small, fixed timestep via numerical integration (e.g., Verlet or leapfrog algorithm [49]). Eventually, this four-step loop yields the position of the atom at a function of time, that is, the trajectory of each atom. In first-principles *ab initio* MD (AIMD) simulations, the system's potential energy is accurately computed using first-principles electron-level methods, and the computational cost of first-principles MD simulations typically scales with the cube of the number of electronic degrees of freedom [19]. The extremely high computation cost would significantly slow down the simulation runtime, and a typical AIMD simulation running on regular computing platforms usually takes a long time from weeks to

months—which limits its potential to virtually screen the entire material design space and discover new materials [21].

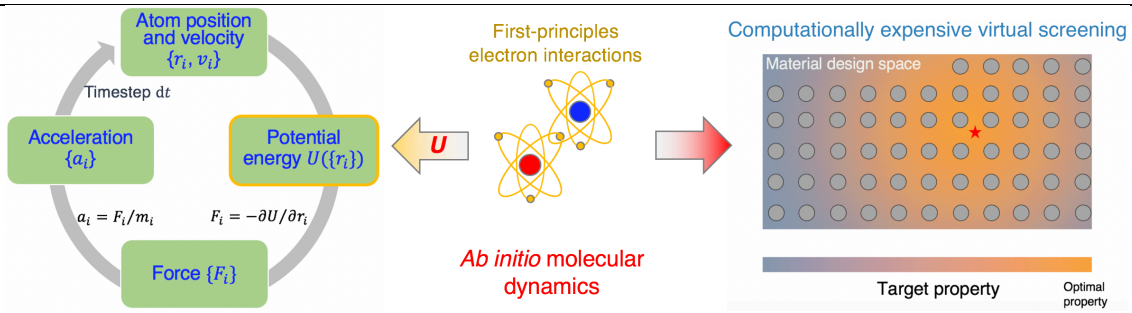


Figure 1-8: Illustration of a computational expensive virtual screening relying on *ab initio* molecular dynamics (AIMD) simulations. AIMD simulation implements Newton's law of motion into a 4-step computational algorithm (see left panel), and the system's potential energy is computed by first-principles formulation of electron-level interactions (see middle panel), which takes numerous computation cost []. In right panel, the color coding represents the target property in the material design space, the red star denotes the optimal material exhibiting optimal property, and the grey circles are the present datapoints explored by AIMD simulations.

1.3.2 Is the Data Informative?

As a supplement to physics-driven modeling, data-driven modeling provides new opportunities to build predictive models for certain materials properties governed by complex, unknown physics [12,46]. However, the accuracy of data-driven modeling, especially machine learning (ML), is largely limited by the non-informativity of datasets in the field of materials science [11].

A successful ML model usually requires large amounts of "informative" data to train the model and produce reasonable predictions. To be considered as "informative", a dataset needs to

be (i) available, (ii) complete, (iii) consistent, (iv) accurate, and (v) representative. However, most materials experiment datasets do not satisfy such requirements, which make the application of ML in materials engineering very challenging [11].

Figure 1-9 shows an example of ML using an uninformative experimental dataset. Since Edisonian "trial-and-error" method has been widely applied in laboratory and industry (see Sec. 1.1.2), the datapoints of experimental datasets are usually chosen with some randomness or slightly adjusted from previous experiments, resulting in sparse data over the entire design space or clustered data in a small region (see Fig. 1-9) [11]. However, both the two types of data curations are not ideal to inform ML models to guide the discovery of optimal material, considering ML models generally excel at interpolating existing datapoints but show poor extrapolatability to ranges away from the existing dataset (see Fig. 1-9) [12]. Overall, the low informativity of materials datasets would reduce the predictivity of data-driven modeling and significantly limits its huge potential for materials discovery.

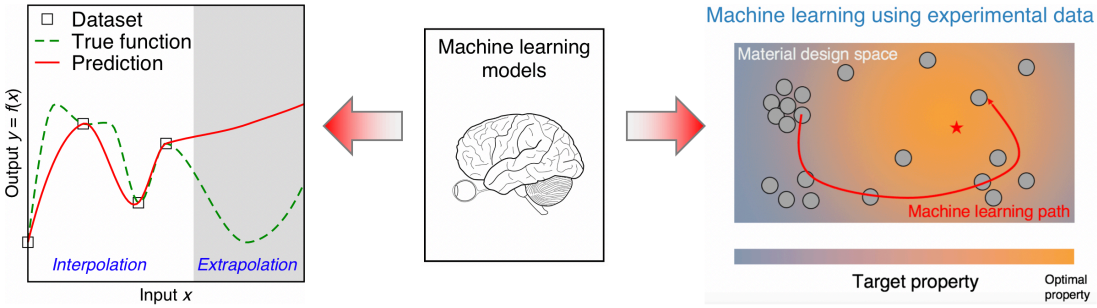


Figure 1-9: Illustration of machine learning (ML) using an uninformative experimental dataset. ML models are good at interpolation but not extrapolation (see left panel). In right panel, the color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by the method, and the red arrow shows the machine learning search path.

1.4 Proposal of Next-Generation Modeling and Thesis Overview

1.4.1 Main Contributions of the Thesis in Next-Generation Modeling

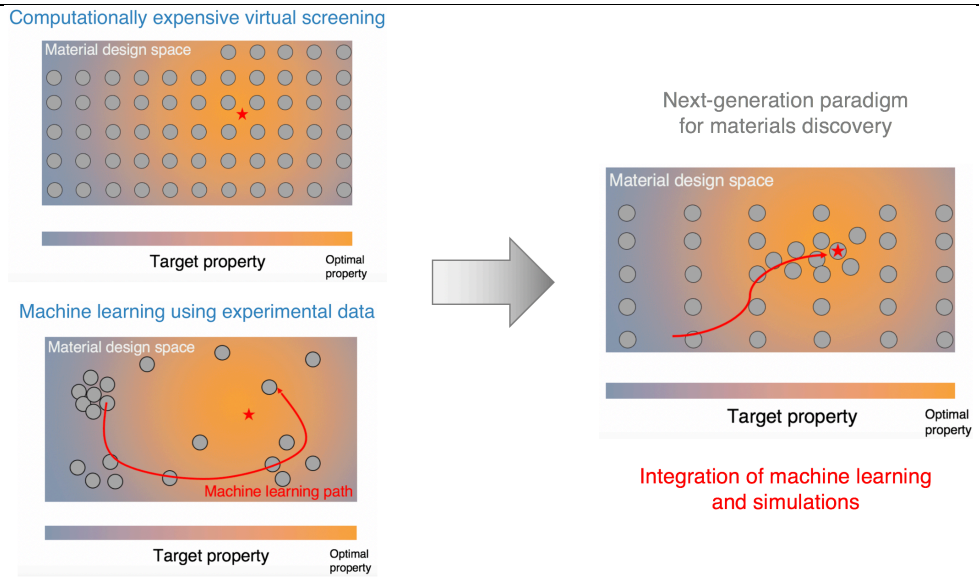


Figure 1-10: Illustration of next-generation paradigm for materials discovery by integrating machine learning and simulations. The color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by the method, and the red arrow shows the machine learning search path.

To address these challenges facing materials modeling—namely, (i) the complex physics and (ii) the uninformative data (see Sec. 1.3), I propose herein the framework of next-generation paradigm for materials discovery, that is, integration of machine learning (ML) and simulations (see Fig. 1-10). On the one hand, computational simulation enables virtual screening of the entire material design space, but the simulations often take numerous computational costs [17]. On the other hand, ML could guide the exploration toward local regions most promising to find optimal material, but the model accuracy is highly dependent on the quality of the dataset [3]. Taking into

account the *pros and cons* of both modeling methods, I envision that the integration of ML and simulations is promising to leverage the predictive power of both modeling methods, and at the same time, overcome the limitations of each individual method (see below). Overall, it is a promise that the predictivity of the integrated modeling is unparallel so as to unlock a new era for materials discovery (see Fig. 1-10) [11,50].

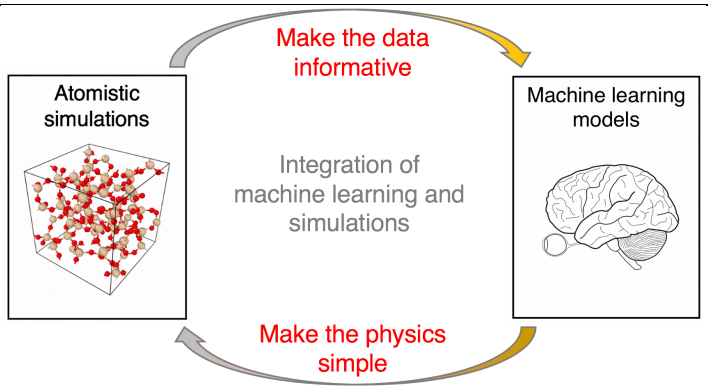


Figure 1-11: Schematic of next-generation modeling via integration of machine learning and simulations. The integrated modeling aims to (i) make the data informative to facilitate data-driven machine learning and (ii) make the physics simple to facilitate physics-driven simulations.

The unparallel predictive power of this integrated modeling method can be understood from two aspects described below (see Fig. 11). First, from the aspect of data-driven ML models, in contrast to traditional experiments, high-throughput simulations are able to produce large amounts of "informative" data to train ML models and provide new opportunities to overcome the limitations using experimental data (see Sec. 1.3.2 and 1.4.3) [11]. Second, from the aspect of physics-driven simulations, since our goal is to facilitate the modeling by adopting less-complex physics laws (see Sec. 1.3.1 and 1.4.2), we expect that, when using ML models to infer the target property landscape in the material design space, the ML models could capture some physics

laws— which are otherwise hard to formulate but govern the property landscape, so as to build simulations driven by machine-learned-physics (see Sec. 1.4.4) [51,52]. Overall, by "fusion" of ML and simulations, this thesis contributes the next-generation materials modeling in leveraging the predictive power of both data-driven ML and physics-driven simulations toward (i) more informative data-driven ML and (ii) less complex physics-driven simulation.

1.4.2 Physics-Driven Modeling: Make the Physics Simple

According to the two grand challenges facing materials modeling (see Sec. 1.3), next-generation materials modeling would contain two types of tasks, namely, (i) make the physics simple to facilitate physics-driven modeling [51,52], and (ii) make the data informative to facilitate data-driven modeling [11,50]. In this section, I would introduce the next-generation modeling driven by simple physics.

To be considered as "simple", physics-driven modeling follows a 3-stage scheme in simplifying the physics principles that govern materials properties, i.e., (i) make the complex (or unknown) physics become interpretable [53], (ii) make the interpretable physics become computational [29], (iii) make the computational physics become computationally efficient [47]. In practice, each stage contains some unique challenges and needs numerous efforts to make some progress. For example, Figure 1-12 shows a computational simplification of *ab initio* MD simulation by classical MD simulation, where *ab initio* MD simulations conduct a computationally expensive calculation of interatomic potentials using the first-principles electron-level methods (see Sec. 1.3.1) [19]. In order to make the computation more efficient, the interatomic potentials can be approximately estimated by some empirical functions, that is, empirical potentials (viz., empirical forcefields) in classical MD simulations (see Fig. 1-12) [47,48]. Nevertheless, this

approximation still needs a lot of consideration about the balance between accuracy and simplicity of empirical forcefields (see Chapter 7–9) [54].

In practice, we have different approaches to fulfill the 3-stage scheme so as to make the physics simple, including experimental [55], theoretical [6], and data-driven analysis [18,56]. This thesis provides examples of all the 3 stages. First, some new types of simulations are developed by using theoretical analysis to make the interpretable physics become computational (see Chapter 2-4) [16,29,57]. Then it is notable that, by integrating machine learning and simulations, this thesis contributes to make the complex physics become interpretable (see Chapter 10-11) [58,59], and make the computational physics become computationally efficient (see Section Chapter 7-9) [24,54,60].

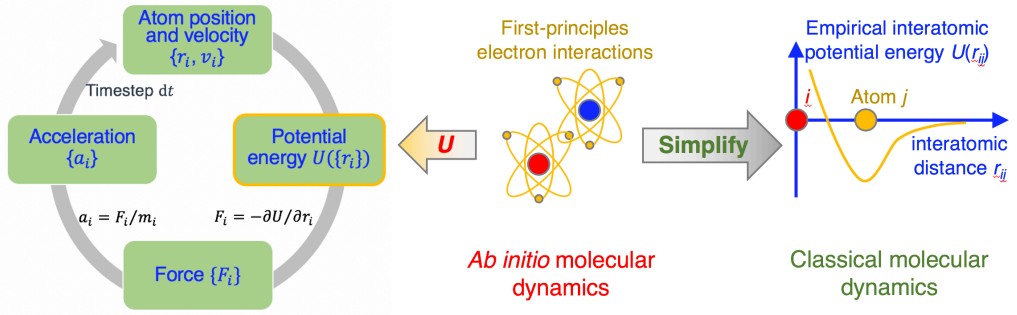


Figure 1-12: Illustration of making the physics simple in molecular dynamics (MD) simulation, where the computationally expensive formulation of first-principles interatomic potentials in *ab initio* MD can be simplified as some empirical potential functionals that take little computation cost in classical MD.

1.4.3 Data-Driven Modeling: Make the Data Informative

Clearly, the bottleneck of next-generation modeling driven by data lies in the quality (viz. informativity) of the dataset used to train the model [61]. As illustrated in Fig 1-13, by making the data more informative—note that, to be considered "informative", a dataset needs to be (i)

available, (ii) complete, (iii) consistent, (iv) accurate, and (v) representative [11]—that is, by reducing the sparseness and randomness of data distribution in the material design space, machine learning models become properly trained to offer more precise guidance in discovering optimal material [3]. In this section, I would introduce the methods that make the data informative toward next-generation data-driven modeling.

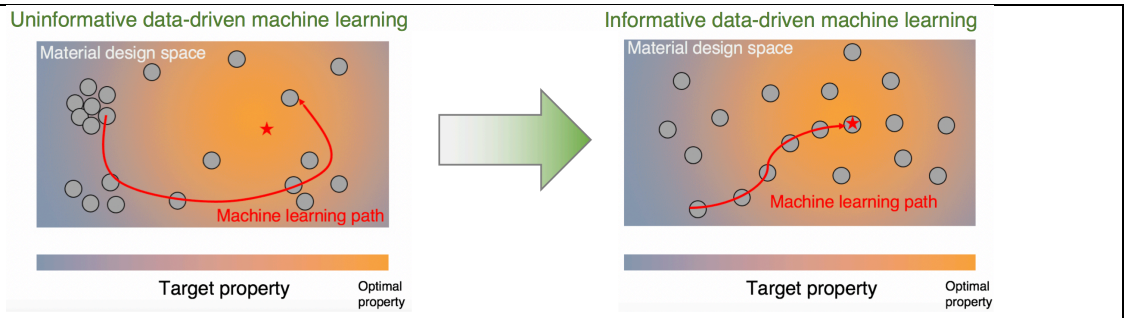


Figure 1-13: Illustration of making the data informative in machine learning, where more informative data-driven machine learning would facilitate materials discovery. The color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by the method, and the red arrow shows the machine learning search path.

As a simple classification, the methods to make the data informative can be divided into two groups: (i) transforming the existing dataset into a more informative format [12], and (ii) adding more "informative" data into the existing dataset [7,8,50]. This thesis contributes to use both types of methods toward more informative data-driven machine learning (ML). First, I use some *prior* physics knowledge to transform an existing dataset into a more informative format to enhance the ML model's extrapolability (see Chapter 5) [12]. Second, by integrating ML and simulations, I enable an automatic addition of informative simulation data in the training of ML

models—which greatly improves the models' prediction accuracy and extrapolability (see Chapter 6) [50].

1.4.4 Fusion of Physics- and Data-Driven Modeling

As both physics- and data-driven modeling have their advantages and bottlenecks individually in next-generation paradigm for materials discovery (see Sec. 1.4.1) [10,18], it become urgently important to "fuse" physics- and data-driven modeling—that is, integration of machine learning and simulations in the scope of this thesis—so as to overcome their individual limitations (see Sec. 1.3) and maximize both modeling tools' predictive power in the integrated pipeline (see Sec. 1.4.1) [11,50].

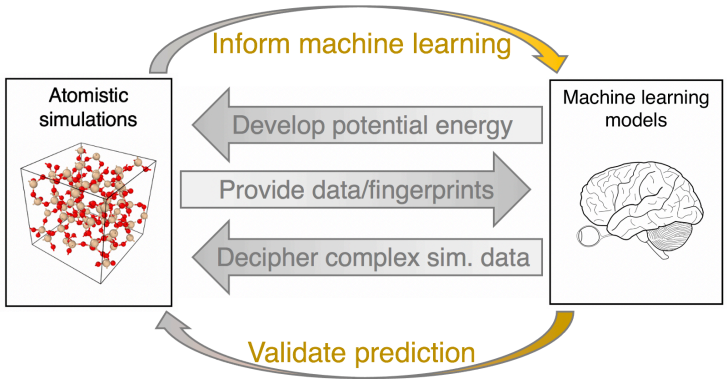


Figure 1-14: Schematic of "fusion" of physics-driven simulations and data-driven machine learning (ML), by (i) providing simulation data/fingerprints to inform ML models [50], (ii) developing potential energy by ML [24], and (iii) deciphering complex simulation data by ML [59].

To demonstrate the predictive power, I investigate several applications of the integrated modeling (see Fig. 15). On the one hand, ML can assist in (i) developing empirical forcefields for accurate and computationally-efficient simulations (Chapter 7-9) [24,54,60], (ii) "separating the

wheat from the chaff" in large amounts of complex simulation data to gain new insights or generate new knowledge of the underlying physics governing materials behaviors (Chapter 10-11) [58,59], and (iii) accelerating simulations by surrogate machine learning engines [62,63]. On the other hand, simulation can generate large amounts of high-fidelity data that can be used to train machine learning models, which, in turn, can be validated by simulations (Chapter 6) [50]. Overall, I envision that smart closed-loop integrations of ML modeling and simulations will leapfrog materials modeling [3,18,36].

1.4.5 Thesis Organization

To address the two grand challenges facing materials modeling (i.e., the complex physics and the uninformative data, see Sec. 1.3), next-generation materials modeling aims to (i) make the physics simple to facilitate physics-driven modeling (see Sec. 1.4.2), and (ii) make the data informative to facilitate data-driven modeling (see Sec. 1.4.3) [3,18,36]. By integrating data-driven machine learning (ML) and physics-driven computational simulations (see Sec. 1.4.4), the next-generation paradigm for materials discovery emerges (see Fig. 1-10), that is, advancing less-complex physics-driven simulations and more-informative data-driven ML models toward unparalleled acceleration for materials discovery [3,18,36].

Based on the interplay between ML and simulation (see Fig. 1-11), this thesis is divided into three sections to describe the methodology of next-generation materials modeling:

- Section A is "*physics-driven computational simulations: make the physics simple*", where I adopted theoretical analysis to make the interpretable physics become computational (see Chapter 2-4) [16,29,57]. Based on the interpretable physics that governs materials behaviors, I simplify the physics laws and formulate the laws into computational models

so as to simulate the material behaviors, including precipitation kinetics (see Chapter 2) [57], nanomechanics (see Chapter 3) [16], and long-time creep deformation (see Chapter 4) [29] of colloidal calcium–silicate–hydrate (C–S–H) gels— the glue of concrete that forms upon the hydration of cement.

- Section B is "*data-driven machine learning: make the data informative*", where I use some *prior* physics knowledge to transform an existing dataset into a more informative format to enhance the ML model's extrapolability (see Chapter 5) [12]. Specifically, in contrast to blind machine learning of glass composition-dissolution rate relationship, I construct some features of the glass network topology as ML input descriptors and apply it to predict the stage I dissolution kinetics (i.e., forward rate, far from saturation) of sodium aluminosilicate glasses (see Chapter 5) [12].
- Section C is "*integration of machine learning and simulations: toward next-generation materials modeling*", where the applications of the integrated modeling can be further divided into three sub-sections:
    - Section C1 is "*toward more informative data-driven machine learning*", where machine learning is informed by differentiable simulations. This enables an automatic addition of informative simulation data in the training of ML models, which greatly improves the models' prediction accuracy and extrapolability (see Chapter 6) [50]. Specifically, taking the example of the inverse design of a porous matrix featuring targeted sorption isotherm, I introduce a deep generative pipeline that combines an end-to-end differentiable simulator with a generator model, and demonstrate the power of this approach in accelerating materials' inverse design (see Chapter 6) [50].

- Section C2 is "*toward less complex physics-driven simulation*", where some accurate-yet-computationally efficient empirical forcefields for classical MD simulations are developed by ML methods (see Chapter 7-9) [24,54,60]. Taking the example of Buckingham-format empirical forcefield for glassy silica, I utilize ML to explore the entire landscape of Buckingham potential to (i) pinpoint the forcefield exhibiting highest accuracy (see Chapter 7) [60], (ii) obtain the balance between forcefield accuracy and simplicity (see Chapter 8) [54], and (iii) compare the competitive forcefields in the "bistability" landscape (see Chapter 9) [24].
- Section C3 is to "*gain new physics knowledge*", where ML can assist in "separating the wheat from the chaff" in large amounts of complex simulation data to gain new insights or generate new knowledge of the underlying physics governing materials behaviors (Chapter 10-11) [58,59], including glass dynamics (Chapter 10) [58], and early-stage creep dynamics of gels (Chapter 11) [59].

1.5 References

[1] J.C. Mauro, C.S. Philip, D.J. Vaughn, M.S. Pambianchi, Glass Science in the United States: Current Status and Future Directions, International Journal of Applied Glass Science. 5 (2014) 2–15. https://doi.org/10.1111/ijag.12058.

[2] J.C. Mauro, E.D. Zanotto, Two Centuries of Glass Research: Historical Trends, Current Status, and Grand Challenges for the Future, International Journal of Applied Glass Science. 5 (2014) 313–327. https://doi.org/10.1111/ijag.12087.

[3] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: A review, Journal of Non-Crystalline Solids: X. 4 (2019) 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[4] E.D. Zanotto, F.A.B. Coutinho, How many non-crystalline solids can be made from all the elements of the periodic table?, Journal of Non-Crystalline Solids. 347 (2004) 285–288. https://doi.org/10.1016/j.jnoncrysol.2004.07.081.

[5] A.K. Varshneya, Fundamentals of Inorganic Glasses, Elsevier, 2013.

[6] K. Binder, W. Kob, Glassy Materials and Disordered Solids: An Introduction to Their Statistical Mechanics, World Scientific, 2011.

[7] T.W. Liao, G. Li, Metaheuristic-based inverse design of materials – A survey, Journal of Materiomics. 6 (2020) 414–430. https://doi.org/10.1016/j.jmat.2020.02.011.

[8] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science. 361 (2018) 360–365. https://doi.org/10.1126/science.aat2663.

[9] R. Jose, S. Ramakrishna, Materials 4.0: Materials big data enabled materials discovery, Applied Materials Today. 10 (2018) 127–132. https://doi.org/10.1016/j.apmt.2017.12.015.

[10] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science, APL Materials. 4 (2016) 053208. https://doi.org/10.1063/1.4946894.

[11] K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, N.M.A. Krishnan, M.M. Smedskjaer, C. Hoover, M. Bauchy, Predicting the Young’s Modulus of Silicate Glasses using High-Throughput Molecular Dynamics Simulations and Machine Learning, Sci Rep. 9 (2019) 1–11. https://doi.org/10.1038/s41598-019-45344-3.

[12] H. Liu, T. Zhang, N.M.A. Krishnan, M.M. Smedskjaer, J.V. Ryan, S. Gin, M. Bauchy, Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning, Npj Mater Degrad. 3 (2019) 1–12. https://doi.org/10.1038/s41529-019-0094-1.

[13] H. Liu, T. Du, N.M.A. Krishnan, H. Li, M. Bauchy, Topological optimization of cementitious binders: Advances and challenges, Cement and Concrete Composites. 101 (2019) 5–14. https://doi.org/10.1016/j.cemconcomp.2018.08.002.

[14] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, Oxford University Press, 2017.

[15] C. Massobrio, ed., Molecular dynamics simulations of disordered materials: from network glasses to phase-change memory allyos, Springer, Cham Heidelberg, 2015.

[16] H. Liu, S. Dong, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Effects of polydispersity and disorder on the mechanical properties of hydrated silicate gels, Journal of the Mechanics and Physics of Solids. 122 (2019) 555–565. https://doi.org/10.1016/j.jmps.2018.10.003.

[17] E.O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery, Annual Review of Materials Research. 45 (2015) 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.

[18] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[19] M. Boero, A. Bouzid, S. Le Roux, B. Ozdamar, C. Massobrio, First-Principles Molecular Dynamics Methods: An Overview, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials, Springer International Publishing, Cham, 2015: pp. 33–55. https://doi.org/10.1007/978-3-319-15675-0_2.

[20] M. Bauchy, Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: A review, Computational Materials Science. 159 (2019) 95–102. https://doi.org/10.1016/j.commatsci.2018.12.004.

[21] H. Liu, Z. Zhao, Q. Zhou, R. Chen, K. Yang, Z. Wang, L. Tang, M. Bauchy, Present Challenges and Future Developments in Atomistic Modeling of Glasses: A Review, Comptes Rendus Geoscience. (2021).

[22] A. Takada, Atomistic Simulations of Glass Structure and Properties, in: P. Richet, R. Conradt, A. Takada, J. Dyon (Eds.), Encyclopedia of Glass Science, Technology, History, and Culture, 1st ed., Wiley, 2021: pp. 221–232. https://doi.org/10.1002/9781118801017.ch2.8.

[23] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Machine Learning Forcefield for Silicate Glasses, ArXiv:1902.03486 [Cond-Mat]. (2019). http://arxiv.org/abs/1902.03486.

[24] H. Liu, Y. Li, Z. Fu, K. Li, M. Bauchy, Exploring the landscape of Buckingham potentials for silica by machine learning: Soft vs hard interatomic forcefields, J. Chem. Phys. 152 (2020) 051101. https://doi.org/10.1063/1.5136041.

[25] F.H. Stillinger, T.A. Weber, Computer simulation of local order in condensed phases of silicon, Physical Review B. 31 (1985) 5262–5271. https://doi.org/10.1103/PhysRevB.31.5262.

[26] M.S. Daw, S.M. Foiles, M.I. Baskes, The embedded-atom method: a review of theory and applications, Materials Science Reports. 9 (1993) 251–310. https://doi.org/10.1016/0920-2307(93)90001-U.

[27] J. Du, Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 157–180. https://doi.org/10.1007/978-3-319-15675-0_7.

[28] L. Huang, J. Kieffer, Challenges in Modeling Mixed Ionic-Covalent Glass Formers, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 87–112. https://doi.org/10.1007/978-3-319-15675-0_4.

[29] H. Liu, S. Dong, N.M.A. Krishnan, E. Masoero, G. Sant, M. Bauchy, Long-term creep deformations in colloidal calcium–silicate–hydrate gels by accelerated aging simulations, Journal of Colloid and Interface Science. 542 (2019) 339–346. https://doi.org/10.1016/j.jcis.2019.02.022.

[30] J.R. Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, 1994. https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

[31] Y. Yu, M. Wang, D. Zhang, B. Wang, G. Sant, M. Bauchy, Stretched Exponential Relaxation of Glasses at Low Temperature, Physical Review Letters. 115 (2015). https://doi.org/10.1103/PhysRevLett.115.165901.

[32] D.J. Lacks, M.J. Osborne, Energy Landscape Picture of Overaging and Rejuvenation in a Sheared Glass, Physical Review Letters. 93 (2004). https://doi.org/10.1103/PhysRevLett.93.255501.

[33] P.G. Debenedetti, F.H. Stillinger, Supercooled liquids and the glass transition, Nature. 410 (2001) 259–267. https://doi.org/10.1038/35065704.

[34] A.D.S. Parmar, M. Ozawa, L. Berthier, Ultrastable Metallic Glasses *In Silico*, Phys. Rev. Lett. 125 (2020) 085505. https://doi.org/10.1103/PhysRevLett.125.085505.

[35] R.L. McGreevy, Reverse Monte Carlo modelling, J. Phys.: Condens. Matter. 13 (2001) R877–R913. https://doi.org/10.1088/0953-8984/13/46/201.

[36] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Accelerating the Design of Functional Glasses through Modeling, Chem. Mater. 28 (2016) 4267–4277. https://doi.org/10.1021/acs.chemmater.6b01054.

[37] M.C. Onbaşlı, A. Tandia, J.C. Mauro, Mechanical and Compositional Design of High-Strength Corning Gorilla® Glass, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling: Applications: Current and Emerging Materials, Springer International Publishing, Cham, 2018: pp. 1–23. https://doi.org/10.1007/978-3-319-50257-1_100-1.

[38] Y.E. Wang, G.-Y. Wei, D. Brooks, Benchmarking TPU, GPU, and CPU Platforms for Deep Learning, (2019). https://arxiv.org/abs/1907.10701v4.

[39] S.J. Russell, P. Norvig, Artificial Intelligence : A Modern Approach, Malaysia; Pearson Education Limited, 2016. http://thuvienso.thanglong.edu.vn/handle/DHTL_123456789/4010.

[40] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ArXiv:1409.1556 [Cs]. (2014). http://arxiv.org/abs/1409.1556.

[41] X. Wu, X. Zhu, G. Wu, W. Ding, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering. 26 (2014) 97–107. https://doi.org/10.1109/TKDE.2013.109.

[42] S. Tsugawa, T. Yatabe, T. Hirose, S. Matsumoto, An Automobile with Artificial Intelligence, in: Proceedings of the 6th International Joint Conference on Artificial Intelligence - Volume 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1979: pp. 893–895. http://dl.acm.org/citation.cfm?id=1623050.1623117.

[43] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[44] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2014.

[45] E.D. Cubuk, R.J.S. Ivancic, S.S. Schoenholz, D.J. Strickland, A. Basu, Z.S. Davidson, J. Fontaine, J.L. Hor, Y.-R. Huang, Y. Jiang, N.C. Keim, K.D. Koshigan, J.A. Lefever, T. Liu, X.-G. Ma, D.J. Magagnosc, E. Morrow, C.P. Ortiz, J.M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K.N. Nordstrom, P.E. Arratia, R.W. Carpick, D.J. Durian, Z. Fakhraai, D.J. Jerolmack, D. Lee, J. Li, R. Riggleman, K.T. Turner, A.G. Yodh, D.S. Gianola, A.J. Liu, Structure-property relationships from universal signatures of plasticity in disordered solids, Science. 358 (2017) 1033–1037. https://doi.org/10.1126/science.aai8830.

[46] S. Gin, J.-M. Delaye, F. Angeli, S. Schuller, Aqueous alteration of silicate glass: state of knowledge and perspectives, Npj Mater Degrad. 5 (2021) 1–20. https://doi.org/10.1038/s41529-021-00190-5.

[47] A. Carré, J. Horbach, S. Ispas, W. Kob, New fitting scheme to obtain effective potential from Car-Parrinello molecular-dynamics simulations: Application to silica, EPL. 82 (2008) 17001. https://doi.org/10.1209/0295-5075/82/17001.

[48] A. Carré, S. Ispas, J. Horbach, W. Kob, Developing empirical potentials from ab initio simulations: The case of amorphous silica, Computational Materials Science. 124 (2016) 323–334. https://doi.org/10.1016/j.commatsci.2016.07.041.

[49] D. Levesque, L. Verlet, Molecular dynamics and time reversibility, J Stat Phys. 72 (1993) 519–537. https://doi.org/10.1007/BF01048022.

[50] H. Liu, Y. Liu, Z. Zhao, M. Bauchy, S.S. Schoenholz, E.D. Cubuk, End-to-End Differentiability and Tensor Processing Unit Computing to Accelerate Materials' Inverse Design, in: 2020.

[51] S. Chmiela, H.E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, Nat Commun. 9 (2018) 3887. https://doi.org/10.1038/s41467-018-06169-2.

[52] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, Nature Materials. 20 (2021) 750–761. https://doi.org/10.1038/s41563-020-0777-6.

[53] E.D. Cubuk, S.S. Schoenholz, J.M. Rieser, B.D. Malone, J. Rottler, D.J. Durian, E. Kaxiras, A.J. Liu, Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods, Physical Review Letters. 114 (2015). https://doi.org/10.1103/PhysRevLett.114.108001.

[54] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Balance between accuracy and simplicity in empirical forcefields for glass modeling: Insights from machine learning, Journal of Non-Crystalline Solids. 515 (2019) 133–142. https://doi.org/10.1016/j.jnoncrysol.2019.04.020.

[55] M. Affatigato, Modern Glass Characterization, John Wiley & Sons, 2015.

[56] J.C. Mauro, Y. Yue, A.J. Ellison, P.K. Gupta, D.C. Allan, Viscosity of glass-forming liquids, Proceedings of the National Academy of Sciences. 106 (2009) 19780–19784. https://doi.org/10.1073/pnas.0911705106.

[57] H. Liu, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Structural percolation controls the precipitation kinetics of colloidal calcium–silicate–hydrate gels, J. Phys. D: Appl. Phys. 52 (2019) 315301. https://doi.org/10.1088/1361-6463/ab217b.

[58] H. Liu, E. Li, E.D. Cubuk, S.S. Schoenholz, S. Xiao, C. Yang, G. Sant, M.M. Smedskjaer, M. Bauchy, Deciphering a Structural Signature of Glass Dynamics by Machine Learning, Physical Review B. (2021).

[59] H. Liu, S. Xiao, L. Tang, E. Bao, E. Li, C. Yang, Z. Zhao, G. Sant, M.M. Smedskjaer, L. Guo, M. Bauchy, Predicting the early-stage creep dynamics of gels from their static structure by machine learning, Acta Materialia. 210 (2021) 116817. https://doi.org/10.1016/j.actamat.2021.116817.

[60] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Parameterization of empirical forcefields for glassy silica using machine learning, MRS Communications. (2019) 1–7. https://doi.org/10.1557/mrc.2019.47.

[61] R. Batra, S. Sankaranarayanan, Machine learning for multi-fidelity scale bridging and dynamical simulations of materials, J. Phys. Mater. (2020). https://doi.org/10.1088/2515-7639/ab8c2d.

[62] D. Kochkov, J.A. Smith, A. Alieva, Q. Wang, M.P. Brenner, S. Hoyer, Machine learning accelerated computational fluid dynamics, ArXiv:2102.01010 [Physics]. (2021). http://arxiv.org/abs/2102.01010.

[63] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, P.W. Battaglia, Learning to Simulate Complex Physics with Graph Networks, ArXiv:2002.09405 [Physics, Stat]. (2020). http://arxiv.org/abs/2002.09405.

Section A. Physics-Driven Computational Simulations: Make the Physics Simple

A1. Grand Canonical Monte Calo Simulation

Chapter 2. Structural Percolation Controls the Precipitation Kinetics of Colloidal Calcium–Silicate–Hydrate Gels

2.1 Introduction

The precipitation process strongly affects the microstructure and properties of colloidal gels [1–3]. This is a consequence of the fact that jammed colloidal gels are out-of-equilibrium—so that they are non-ergodic and their properties depend on their history [2]. During precipitation, microstructure (i.e., the spatial organization of the colloidal grains) and kinetics (i.e., precipitation rate) are closely related to each other [2–4]. Based on these linkages, the precipitation kinetics can be tuned by different experimental methods, such as chemical modification of the grain surfaces, control of the grain concentration, or reaction temperature and pressure [2]. In turn, the precipitation kinetics eventually controls the microstructure of the final jammed colloidal gel [3,4]. However, little remains known about the mutual linkages between microstructure and precipitation kinetics.

Calcium–silicate–hydrate (C–S–H) gel—the glue of concrete that forms upon the hydration of cement—is a technologically-important inorganic colloidal hydrogel material [5,6]. Importantly, the colloidal microstructure of C–S–H largely controls its mechanical response and governs concrete strength [7–9].This is significant as, due to the large carbon impact of cement and concrete, improving the mechanical properties of C–S–H would permit to use less material while achieving constant performance [3,10,11]. Hence, enhancing the properties of C–S–H largely relies on our ability to finely tune the colloidal structure of C–S–H during and after the precipitation process. However, the complex and heterogeneous nature of cement pastes largely limit our ability to experimentally characterize the time-dependent microstructure of C–S–H.

Recently, based on coarse-grained mesoscale simulations, Masoero *et al.* introduced a colloidal model that offers a realistic description of the mesoscale structure and nanomechanics of C–S–H [6]. Ioannidou e*t al.* also proposed an alternative model to describe the early-stage gelation of C–S–H [4]. These colloidal models have permitted to investigate the relationship between structure and precipitation kinetics [3,4,12]. However, several questions remain unanswered. What are the structural features that govern the precipitation kinetics of the gel? In turn, how does the kinetics of precipitation control the microstructure of the gel at setting? Answering these questions would facilitate the design of new gel formulations that can set "on demand" at low or high packing density.

Here, based on grand canonical Monte Carlo (GCMC) simulations, we investigate the precipitation mechanism of C–S–H gels. We show that both the thermodynamics and kinetics of the precipitation of C–S–H are governed by the underlying percolation of its microstructure. Further, we demonstrate that the critical gel packing density at which percolation occurs is controlled by the balance between the size and shape of the globular clusters that form upon precipitation. These results highlight the retroactive nature of the linkages between structure and precipitation dynamics.

This paper is organized as follows. In Sec. 2.2, we describe the simulation methodology used to model and analyze the precipitation of C–S–H. In Sec. 2.3, we investigate the effect of structural percolation of the C–S–H grains on the kinetics and thermodynamics of precipitation. In turn, the effect of the precipitation kinetics on the nature of the structural percolation of C–S–H is discussed in Sec. 2.4. Finally, some conclusions are presented in Sec. 2.5.

2.2 Methods

2.2.1 Inter-grain interactions

We adopt here the coarse-grained mesoscale model of C–S–H introduced by Masoero *et al.* [6], which has been shown to offer an excellent description of C–S–H structure and mechanical properties [9,13,14]. The C–S–H gel is here described as an ensemble of monodisperse C–S–H grains with a diameter of 5 nm. The grains interact with each other through a generalized Lennard-Jones interaction energy potential:

$$U_{ij}(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{2\alpha} - \left(\frac{\sigma}{r_{ij}}\right)^{\alpha}\right] \quad \textbf{Eq. (2-1)}$$

where σ is the grain diameter (5 nm here), α a parameter that controls the narrowness of the potential well, r_{ij} the distance between the centers of a pair of grains i and j, and ε the depth of the potential energy well. By considering each pair of grains in contact as two springs in series, the energy depth is given by $\varepsilon = A_0\sigma^3$, where $A_0 = kE$ is a prefactor that is proportional to the Young's modulus E of a bulk C–S–H grain ($E = 63.6$ GPa based on previous atomistic simulations of bulk C–S–H [15]) and $k = 0.002324$ (computed from a serial spring model [9]). The potential defined in Eq. (2-1) exhibits a minimum at $r_m = \sqrt[\alpha]{2}\sigma$ so that the effective diameter of a grain i is here defined as $\sigma_0 = \sqrt[\alpha]{2}\sigma$. The attractive force shows a maximum at the distance $r_u = \sqrt[\alpha]{\frac{4\alpha+2}{\alpha+1}}\sigma$ so that, by choosing $\alpha = 14$, the tensile strain at failure $\varepsilon_u = (r_u - r_m)/r_m$ is close to the value of 5% obtained in previous atomistic simulation of bulk C–S–H [9,15,16].

2.2.2 Grand Canonical Monte Carlo simulations

The C–S–H gel configurations are generated by grand canonical Monte Carlo (GCMC) simulations using the open-source LAMMPS package [17], as described in the following. Starting

from an initially empty cubic box of length 600 Å with periodic boundary conditions, C–S–H grains are iteratively inserted to mimic the precipitation process [4]. Each GCMC step comprises *X* attempts of grain insertions or deletions followed by *Y* attempts to randomly displace an existing grain. At each step, the probability of acceptance of the attempt is given by $\min\{1, \exp[-(\Delta U - \mu\lambda)/k_{\mathrm{B}}T]\}$, where k_{B} is the Boltzmann constant, T the temperature, ΔU the variation in potential energy caused by the trial move, μ the chemical potential (taken here as $-2k_{\mathrm{B}}T$ based on previous studies [4,12]), and λ the variation in the number of C–S–H grains [18]. This GCMC step is then iteratively repeated until the number of inserted grains reaches a plateau. During the precipitation process, the packing fraction φ of each gel configuration is computed as $\varphi = n_g \pi \sigma^3/6V$, where n_g is the number of inserted grains at a given time and V is the volume of the simulation box. Based on the system size considered herein, we typically get $n_g \approx 1700$ at saturation.

The extent of structural relaxation in the gel upon precipitation strongly depends on the number of grain displacements in between two insertions/deletions. To quantify this effect, we define here a kinetic rate *R* as:

$$R = X/Y \qquad \textbf{Eq. (2-2)}$$

R is a metric that is qualitatively equivalent to a precipitation rate as it characterizes the inverse of the duration during which the grains are allowed to reorganize in between two successive insertions. Namely, a large *R* value corresponds to a high precipitation rate, wherein the grains have only limited opportunity to reorganize during precipitation. Here, to investigate the effect of the kinetics of precipitation, six values of the kinetic rate *R* (1.0, 0.5, 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}) are considered.

The precipitation dynamics is dominated by not only the kinetic rate, but also the diffusion time of the system, i.e., the characteristic time that is needed for the grains to overcome some activation energy barriers. Although our GCMC simulations do not rely on any explicit time, it has been suggested that the "real" precipitation time τ should be proportional to the logarithm of the number of simulation steps [4]:

$$\tau \propto \log(N) \qquad \textbf{Eq. (2-3)}$$

where N is the total number of steps insertions/deletions or displacements, i.e., $N = n \times (X + Y)$ after n GCMC steps. This can be understood from the fact that, although it is instantaneous in our simulations, the precipitation of a C–S–H grain would occur within a typical timescale τ_0, which increases as the system becomes more and more packed. Assuming that, to the first order, $\tau_0 \propto N$, an increment in real time $\Delta\tau$ should be compared with an increment $\Delta N/N$ (or $\Delta N/\tau_0$) in our simulations, which yields Eq. (2-3) by integration [4].

2.2.3 Analysis of the clusters of C–S–H grains

To track the structural percolation of the C–S–H configuration upon the addition of grains in the simulation box, we compute the number, size, and shape of the clusters of C–S–H grains. Two C–S–H grains are here defined as belonging to the same cluster if their distance is lower than 6 nm—note that this cutoff distance is here defined as the position of the first minimum after the main peak in the pair distribution function, that is, the extent of the first coordination shell. Periodic boundary conditions are considered in the determination of each cluster. We then compute the total number of clusters c and the average number of grains per cluster g. The length of the largest cluster L_{max} in a given C–S–H configuration is calculated as follows. The dimensions L_x, L_y, and L_z of each cluster along the x, y, and z axis are first calculated by taking into account the periodic

boundary conditions. $L_{\max}$ is then defined as the largest dimension of any cluster among all the clusters in this ensemble. Note that we only consider here the projections of the cluster dimensions on the Cartesian axis as we aim to compare these dimensions to the size of the simulation box. The normalized length of the largest cluster is then defined by normalizing this quantity by the length of the cubic simulation box L_{box}:

$$L = L_{\max}/L_{\mathrm{box}} \qquad \textbf{Eq. (2-4)}$$

To characterize the shape of the clusters, we then compute an "aspect ratio" parameter *A* for the clusters as detailed in the following. First, for each individual cluster, the pair of grains belonging to the cluster and exhibiting the largest distance from each other is identified. The positions of these two grains are then used to define the diameter of a sphere that, by construction, entirely contains the cluster. The shape of the cluster is then characterized by randomly inserting some points within the sphere. Each point is defined as being part of the cluster if its distance from the center of the nearest grain is less than 2 times the grain diameter. Periodic boundary conditions are taken into account throughout the process. The aspect ratio parameter of the cluster is then defined as the fraction of these points that is part of the cluster. Note that, for a perfectly spherical cluster, this aspect ratio would be equal to 1 since all the randomly inserted points would be part of the cluster. In contrast, more elongated clusters are characterized by A < 1. To exclude the effect of isolated grains or tiny clusters, only the ones that are made of more than 5 grains are considered for this calculation. To filter out the effect of statistical fluctuations, the clustering analysis is performed for six independent simulations of precipitation simulations for each kinetic rate. All the clustering results presented thereafter are averaged over these configurations.

2.2.4 Tracking the percolation of the microstructure

Following classical percolation theory, we describe the evolution of normalized length of the largest cluster L as a function of the packing fraction φ as [19,20]:

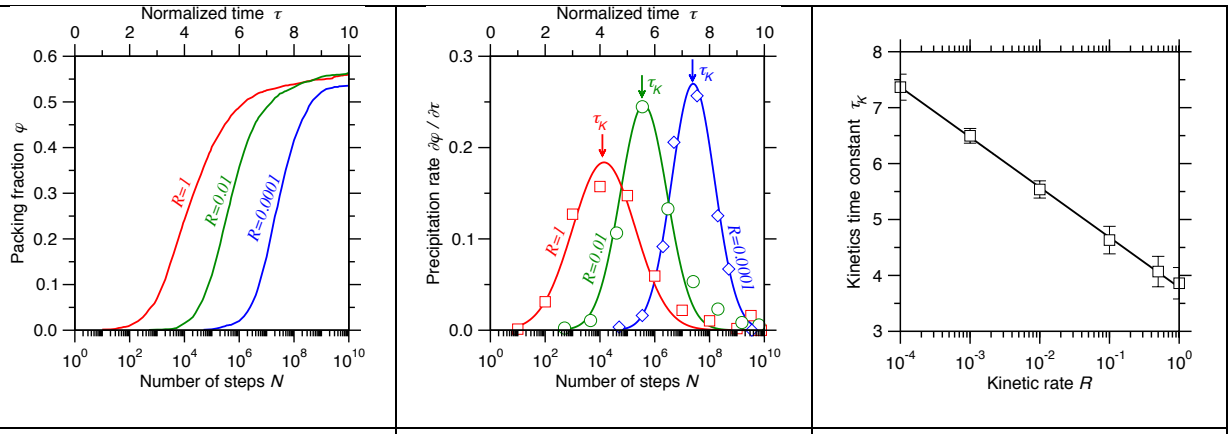
$$L \propto (\varphi_c - \varphi)^{-\nu} \quad \textbf{Eq. (2-5)}$$

where φ_c is the critical percolation threshold, i.e., the packing fraction at which percolation occurs, and ν is the percolation exponent. Namely, when φ approaches φ_c, L tends to infinity and the system is percolated. Here, φ_c was determined as the packing fraction for which the length of the largest cluster reaches the dimension of the simulation box, i.e., $L_{\max} = L_{\text{box}}$ or $L = 100\%$. The critical percolation exponent ν is then determined by fitting the slope of the logarithm of L as a function of $(\varphi_c - \varphi)$. This analysis is conducted on the independent simulations of precipitation performed for each kinetic rate to calculate the mean and standard deviation of φ_c and ν.

2.3 Results

2.3.1 Kinetics of C–S–H precipitation

We first analyze the kinetics of the precipitation of C–S–H and its relationship to the kinetic rate R (see Eq. (2-2)). Figure 2-1(a) shows the evolution of the C–S–H packing density φ as a function of time. Overall, we observe that φ increases monotonically with time and exhibit the typical sigmoidal shape that is observed experimentally [4]. In the following, the precipitation rate of C–S–H is defined as $\partial\varphi/\partial\tau$, that is, the increase in the packing density resulting from the successful insertions of C–S–H grains per unit of time. As shown in Fig. 2-1(b), C–S–H's precipitation kinetics exhibits an initial acceleration stage, which is followed by a deceleration—in agreement with experimental observations [4].



(a) **(b)** **(c)**

Figure 2-1: (a) Packing fraction of the C–S–H gel and **(b)** precipitation rate (slope of the packing fraction) as a function of the number of simulation steps and normalized time for select kinetic rates R—see Eq. (2-2). (c) Kinetics time constant τ_K (i.e., time at which the precipitation rate is maximum) as a function of the kinetic rate R.

We observe that the final packing density at saturation is around 0.6—that is, close to the packing density of random monodisperse spheres [21,22]—and does not significantly depend on the kinetic rate. In contrast, C–S–H precipitation kinetics is found to be largely affected by the the kinetic rate imposed in the simulations (see Figs. 2-1(a) and 2-1(b)). As expected, we observe that a lower kinetic rate results in delayed precipitation. To further quantify the relationship between precipitation kinetics and kinetic rate, we define here a kinetic time constant τ_K as the time at which the maximum precipitation rate occurs, that is, at the transition between the accelerating and decelerating regime. We find that τ_K decreases logarithmically with increasing kinetic rate (see Fig. 2-1(c)), which *a posteriori* justifies the assumptions used to establish Eq. (2-3)—that is, that the real time should be compared with the logarithm of the simulation time lapse.

Finally, we observe that the maximum in the precipitation rate of C–S–H becomes more and more sharp with decreasing kinetic rate (see Fig. 2-2(b)). This suggests that, as the

precipitation dynamics rate decreases, the range of packing density values over which the jamming transition occurs in the gel tends to become narrower. This echoes the fact that, in glasses, the glass transition typically becomes more and more well-defined (that is, it occurs over a narrower range of temperature values) upon decreasing cooling rate [23,24].

2.3.2 Thermodynamics of C–S–H precipitation

We now investigate the thermodynamics of the precipitation of C–S–H, with a focus on potential energy and pressure. Figure 2-2(a) shows the time-evolution of the potential energy of the system. As expected, we find that the potential energy of the system decreases with time, since more and more C–S–H grains are inserted within the box and start interacting with each other. Further, we observe that the trend of the potential energy closely follows that of the packing fraction (see Fig. 2-1(a)), namely, a lower kinetic rate results in a delayed decrease of potential energy, whereas the final of the potential energy of the system at saturation does not significantly depend on the kinetic rate.

Since the potential energy captures the level of cohesion of the system, the results presented in Fig. 2-2 highlight a transition from a non-cohesive liquid state (at low packing density) to a cohesive "packed" state (at high packing density) [25]. This arises from the fact that the system initially consists of isolated, non-interacting C–S–H grains (or clusters of grains) and eventually becomes cohesive as it approaches the jamming transition (see below). To further characterize the kinetics of this transition, we define here an energy time constant τ_U as the time at which the potential energy exhibits the highest concave curvature, that is, when $\partial^2 U/\partial \tau^2$ is minimum (see Fig. 2-2(b))—which captures the time at which the potential starts to drop and becomes non-trivial. As shown in Fig. 2-2(c), we find that τ_U logarithmically decreases with increasing kinetic rate.

This highlights the fact that the thermodynamics of C–S–H during precipitation is strongly affected by the precipitation kinetics.

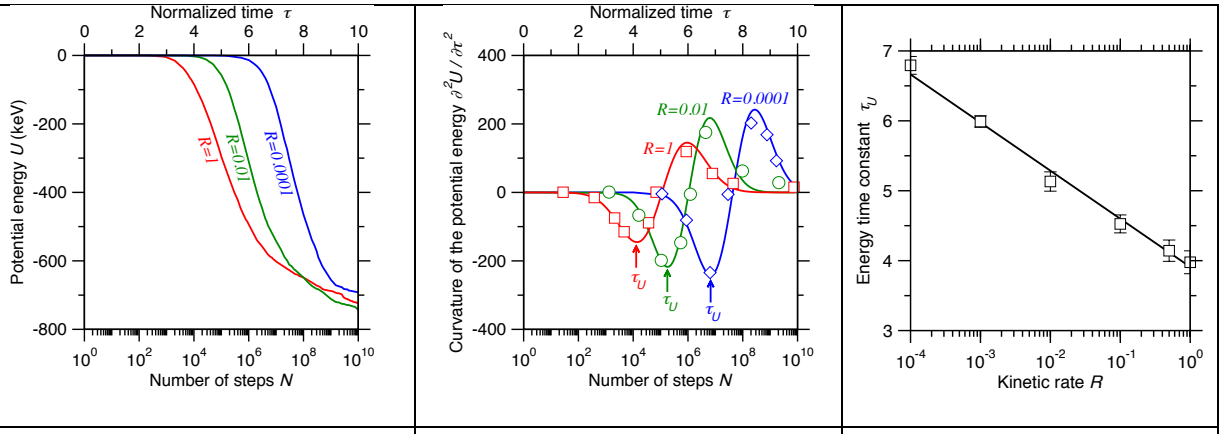


Figure 2-2: (a) Potential energy of the system and **(b)** curvature thereof (i.e., second derivative of the potential energy) as a function of the number of simulation steps and normalized time for select kinetic rates R. **(c)** Energy time constant τ_U (defined as the time at which the potential energy exhibits the highest concave curvature) as a function of the kinetic rate R.

We now focus on the relationship between precipitation kinetics and pressure. Fig. 2-3(a) shows the time-evolution of the pressure undergone by the gel upon precipitation. The system pressure is calulcated by the force virial for all pairwise interactions [26]. We note that evolution of pressure closely follows those of the packing density (see Fig. 2-1(a)) and potential energy (see Fig. 2-2(a)). In details, we observe that, initially, the system is at neglectible pressure, which arises from the fact that the systems comprises isolated, weakly-interacting grains. However, at some point, the pressure undergone by the gel starts to decrease and becomes negative (which is indicative here of a state of tension). This arises from the fact that the precipitation of C–S–H occurs in isochoric conditions—so that, upon precipitation, the grains are attracted to each other

and the system would shrink if the volume was free to change. As such, the onset of such stress is a direct manifestation of the colloidal (rather than granular) nature of C–S–H, which results from the existence of attractive forces among grains. Note that such tensile stress is also observed experimentally [25] as C–S–H typically precipitates in confined conditions. The tensile stress that forms in C–S–H during precipitation largely contributes to the physical shrinkage of cement pastes over time [25].

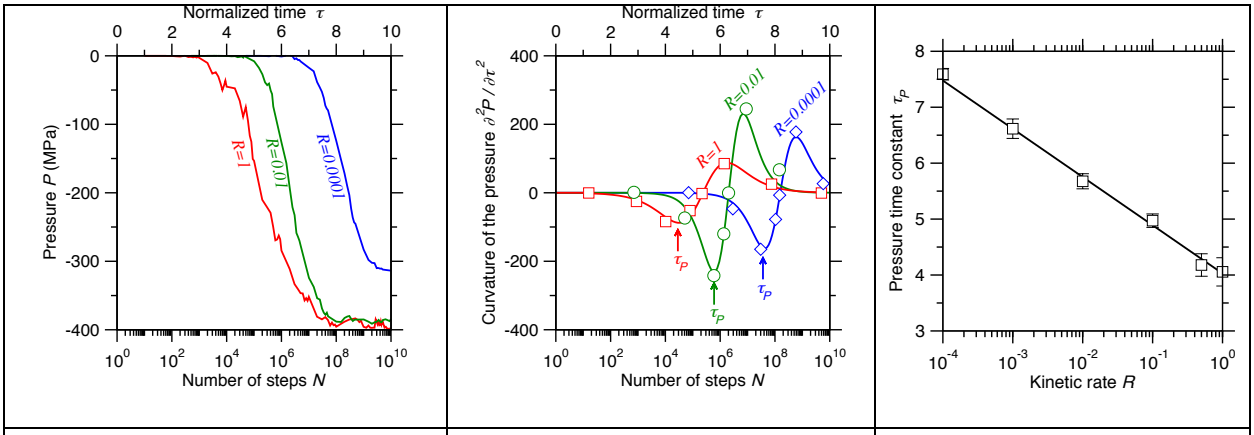


Figure 2-3: (a) Pressure (a negative value being here indicative of a state of tension) and **(b)** curvature thereof (i.e., second derivative of the pressure) as a function of the number of simulation steps and normalized time for select kinetic rates R. **(c)** Pressure time constant τ_P (defined as the time at which the pressure exhibits the highest concave curvature) as a function of the kinetic rate R.

To further characterize the relationship between precipitation kinetics and onset of stress, we define here a pressure time constant τ_P as the time at which the pressure exhibits the highest concave curvature, that is, when $\partial^2 P/\partial\tau^2$ is minimum (see Fig. 2-3(b)). This time constant captures the time at which the pressure starts to drop and becomes non-trivial. As shown in Fig. 2-3(c), we find that τ_P decreases logarithmically with the kinetic rate. This highlights once again the

fact that the thermodynamics of C–S–H during precipitation is strongly controlled by the precipitation kinetics.

Finally, we note that, in contrast with the cases of the potential energy or packing fraction, the final pressure at saturation exhibits a stronger dependence on the kinetic rate (see Fig. 2-3(a)) as the final pressure at saturation decreases with lower values of the kinetic rate. This indicates that the level of tensile stress that forms in C–S–H does not only depend on the number of grains inserted within the system, but also on the overall mesostructure of the gel. This likely arises from the fact that, upon slower precipitation, the C–S–H gel is able to further relax in between the insertation of each grain (see Sec. 2.4), thereby limiting the formation of internal stress [27,28]. This contrasts with the potential energy of the system, which appears to be less sensitive to the structure of the gel.

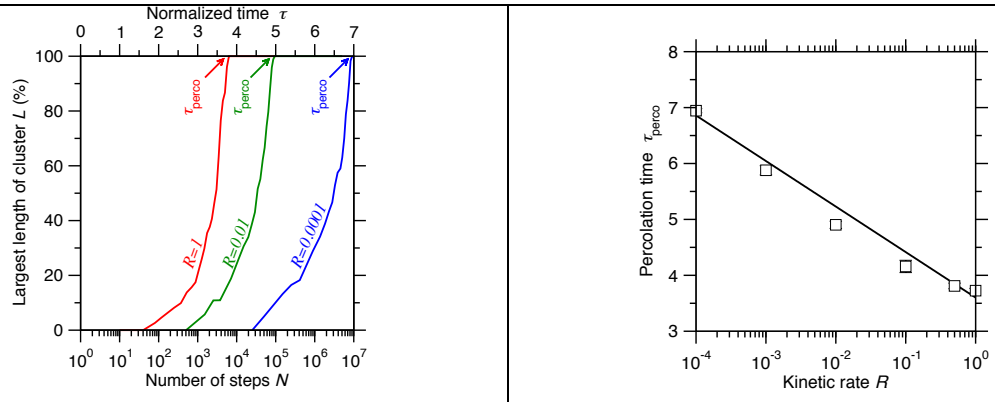
2.3.3 Role of the structural precipitation in the gel

Finally, we investigate how the dynamics and thermodynamics of the precipitation of the C–S–H gel are controlled by its underlying mesostructure. We observe that, starting from an initial situation wherein the C–S–H grains form some isolated clusters due to limited structural reorganization, the mesostructure eventually becomes fully-connected as precipitation proceeds. This picture suggests that the precipitation of C–S–H is associated with a percolation of its mesoscale structure, as detailed in the following.

Fig. 2-4(a) shows the time-evolution of the length L of the largest C–S–H cluster, normalized by the size of the simulation box (see Sec. 2.2.3). As expected, we observe that L increases over time as the mesostructure becomes more and more interconnected, until the length of the largest cluster becomes equal to the size of the simulation box (i.e., $L_{max} = L_{box}$ or $L =$

100%)—which is indicative of percolation. Further, we observe that the evolution of L can be well described by the framework of classical percolation theory, namely, L follows a power-law dependence with respect to the packing density of the gel (see Eq. (2-5) and Sec. 2.2.4) [19,20].

To further describe the relationship between precipitation kinetics and structural percolation, we define a percolation time τ_{perco} as the time at which percolation occurs (i.e., L_{max} = L_{box}). As shown in Fig. 2-4(b), we observe that τ_{perco} decreases logarithmically with the kinetic rate. This suggests that precipitation kinetics and structural percolation are closely related to each other, as detailed in the following.



(a) **(b)**

Figure 2-4: (a) Length of the largest C–S–H cluster normalized by the size of the simulation box (L, see Eq. (2-4)) as a function of the number of simulation steps and normalized time for select kinetic rates R (wherein a length of 100% indicates that the system exhibits percolation). **(b)** Percolation time τ_{perco} (time at which the system becomes percolated, i.e., L = 100%) as a function of the kinetic rate R.

We now further compare the time-evolution of the precipitation kinetics (τ_K), thermodynamics (τ_U and τ_P), and structural percolation (τ_{perco}) of the C–S–H gel (see Fig. 2-5).

Interestingly, we observe that the typical times associated with the (i) precipitation kinetics (i.e., time at which the accelerating-to-decelerating transition occurs) and (ii) thermodynamics (i.e., times at which the system becomes cohesive and at which some tensile stress forms) are both largely correlated to time at which the mesostructure of C–S–H features percolation.

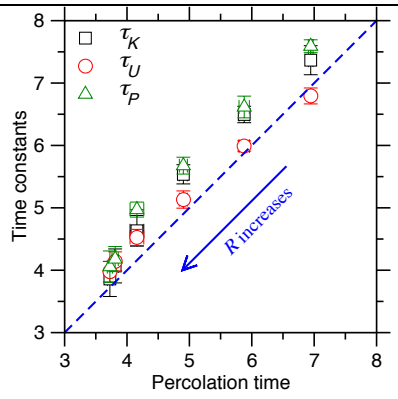


Figure 2-5: Time constants associated with the precipitation kinetics τ_K (i.e., time at which the precipitation rate is maximum), energy τ_U (i.e., time at which the potential energy exhibits the highest concave curvature), and pressure τ_P (i.e., time at which the pressure exhibits the highest concave curvature) as a function of the time constant associated with structural percolation τ_{perco} (i.e., time at which the mesostructure of C–S–H becomes percolated).

On the one hand, the relationship between precipitation kinetics and structural percolation identified herein suggests that the transition from the accelerating to the decelerating regime is closely controlled by the underlying percolation of the mesostructure of C–S–H. This can be understood from the fact that, at low packing density, the C–S–H cluster can freely grow. At this stage, the growth rate is roughly proportional to the external surface of the C–S–H and, thereby, increases over time as the C–S–H clusters grow. In contrast, as percolation occurs, the ease of inserting new C–S–H grains starts to decrease, which results in a decrease in the precipitation kinetics.

On the other hand, the relationship between thermodynamics and structural percolation identified herein suggests that the degree of cohesion and internal stress present in the colloidal structure of the C–S–H gel are also closely correlated to the percolation of its mesostructure. This can be understood from the fact that the degree of cohesion of C–S–H should indeed largely depend on the size of the grain clusters. In turn, it suggests that, starting from an initial situation wherein the C–S–H clusters are mutually isolated and, thereby, do not impose any macroscopic stress, the time at which some internal stress forms within the system corresponds to the time at which the structure percolates—so that stress also percolates through the simulation box.

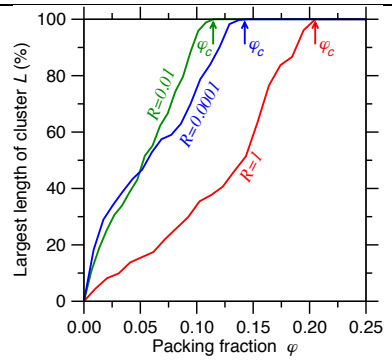
More generally, these results highlight the close correlation between the structure, kinetics, and thermodynamics of C–S–H upon precipitation. Nevertheless, we observe the difference between the thermodynamics, kinetics, and structural time constants tends to slightly increase with lower kinetic rate values (see Fig. 2-5). This suggests that, as the precipitation kinetics decreases, some level of self-organization may occur due to more relaxation time within the C–S–H structure—for instance, to limit the onset of internal stress (see Fig. 2-3(a)). This echoes the "intermediate phase" that has been reported to form in optimally-connected isostatic structural glasses, wherein the atomic network self-organizes to become rigid while avoiding the formation of any internal stress [29–33].

2.4 Discussion

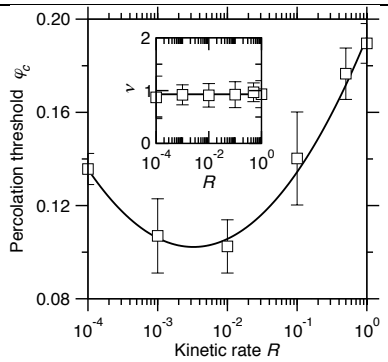
2.4.1 Gel packing density at percolation

Having established that the percolation of the C–S–H mesostructure is closely correlated to its precipitation kinetics, we now discuss how, in turn, the precipitation kinetics controls the structure of C–S–H. First, we note that the evolution of the length of the largest cluster as a function

of packing density can be well described within the framework of classical percolation theory [19,20] (see Fig. 2-6(a) and Sec. 2.2.4). This allows us to extract the critical packing density φ_c at which the percolation occurs and percolation exponent ν (see Eq. (2-5)). We observe that ν does not significantly depend on the kinetic rate and remains close to 0.9. This value matches with the one expected in the case of random-site percolation [19]. In contrast, interestingly, we observe that φ_c exhibits a non-monotonic dependence on the kinetic rate (see Fig. 2-6(b)) as φ_c features a minimum for intermediate values of kinetic rate and increases for low and high kinetic rates. This signals that the kinetics of precipitation significantly affects the nature of the structural percolation occurring in C–S–H, as discussed in the following. This also indicates that tuning the precipitation rate can be used to as an efficient method to alter the degree of packing of the C–S–H gel during setting [34,35].



(a)



(b)

Figure 2-6: (a) Length of the largest C–S–H cluster, normalized by the size of the simulation box, as a function of the packing fraction for select kinetic rates R. **(b)** Percolation critical threshold φ_c (packing fraction at which the system becomes percolated) as a function of the kinetic rate R. The inset shows the percolation exponent ν (see Eq. (2-5)) as a function of R.

2.4.2 Evolution of the gel microstructure upon percolation

Finally, we investigate how the mesostructure of C–S–H is controlled by the precipitation kinetics—with a focus on understanding the structural origin of the non-monotonic evolution of the critical packing density at percolation (see Fig. 2-6(b)). Some illustrative snapshots of C–S–H mesostructures at percolation are presented in Fig. 2-7. On the one hand, we observe that, at large kinetic rate (i.e., large precipitation rate), the structure of C–S–H comprises a large number of small isolated clusters. In contrast, at lower kinetic rate (i.e., slower precipitation rate), the C–S–H structure comprises fewer, but larger clusters. On the other hand, we note that the overall shape of the cluster also depends on the kinetic rate, as large spherical and compact clusters tend to form at low kinetic rate (see Fig. 2-7). These observations suggest that the non-monotonic evolution of the critical packing density at percolation arises from a balance between the size (i.e., number of grains in cluster) and shape (i.e., sphericity and openness of cluster structure) of the C–S–H clusters forming upon percolation.

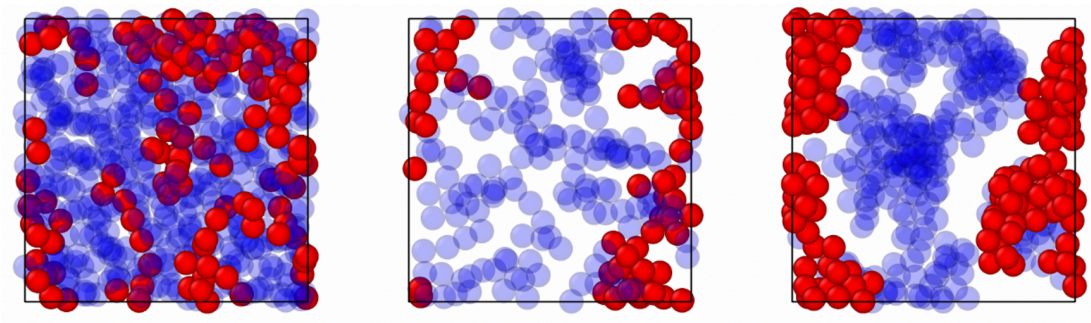


Figure 2-7: Snapshots of colloidal C–S–H structures at the percolation threshold (i.e., $\varphi = \varphi_c$) for select kinetic rates R. In each case, the largest cluster (red) is differentiated from the other clusters (blue).

To establish this mechanism, Fig. 2-8 quantifies the number, size, and shape of the C–S–H clusters at the percolation threshold (i.e., $\varphi = \varphi_c$). First, we note that, at percolation, the number of clusters increases with increasing kinetic rate (see Fig. 2-8(a)), whereas the average size of the cluster decreases (see Fig. 2-8(b)). These results demonstrate that low precipitation rates tend to favor the formation of few large clusters, as the grains have the ability to significantly move and agglomerate in between two successive insertions. In contrast, high precipitation rates result in the formation of a large number of isolated clusters. This scenario is further supported by the fact that, upon decreasing kinetic rate, (i) larger clusters gradually become dominant and (ii) the average coordination number of the grains increases as the clusters become more compact. This can be understood as a "Tetris effect" [3], wherein precipitation is so fast that the C–S–H grains do not have enough time to agglomerate. The lack of grain agglomeration at high kinetic rate explains why percolation is delayed and occurs at larger packing density.

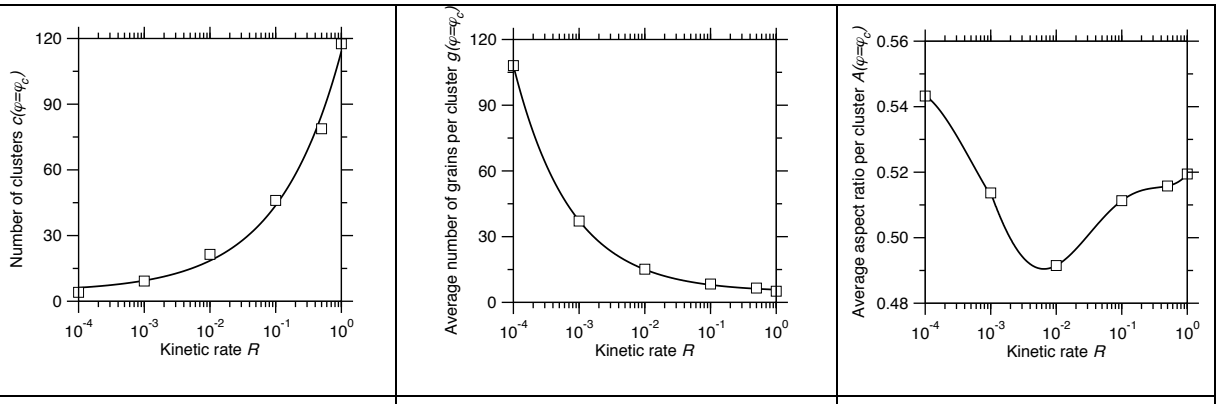


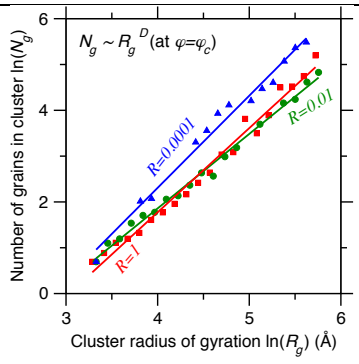
Figure 2-8: (a) Number of clusters c, **(b)** average number of grains per cluster g, and **(c)** average aspect ratio of the clusters A at the percolation threshold (i.e., $\varphi = \varphi_c$) as a function of kinetic rate R. Lower aspect ratio values are indicative of more elongated (i.e., less spherical) clusters.

Finally, although low kinetic rate values tend to favor the formation of larger clusters, Fig. 2-8c shows that such clusters tend to be more spherical (i.e., an aspect ratio that is closer to 1) at low precipitation rate. This can be understood from the fact that, at low precipitation rate, the grains can freely reorganize to form spherical clusters—i.e., to minimize surface energy effect. In contrast, intermediate values of kinetic rates yield more elongated clusters (i.e., lower aspect ratio). Such elongated clusters are more likely to induce structural percolation at low packing density. This explains why percolation tends to occur at larger packing density values at very low kinetic rate. This confirms that the minimum of the critical packing density at which C–S–H exhibits structural percolation arises from the formation of C–S–H clusters that are simultaneously large (in terms of average number of grains) and elongated (i.e., non-spherical). To further elucidate the origin of this minimum, we calculate the fractal dimension D of the structure at the percolation threshold for each kinetic rate (see Fig. 2-9) by using the following equation [36]:

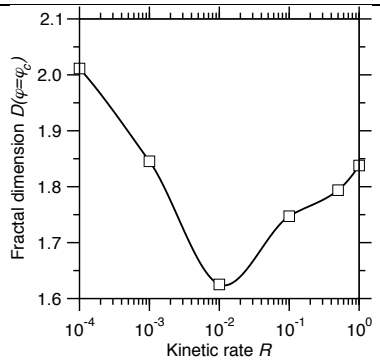
$$N_g \propto {R_g}^D \qquad \textbf{Eq. (2-6)}$$

where N_g is the number of grains in each cluster and R_g is the radius of gyration of the cluster. Note that a fully compact structure exhibits $D = 3$, while D tends to decrease in the case of more open structures. As shown in Fig. 2-9(b), we find that the dimensionality D shows a minimum at intermediate kinetic rate. This strongly supports the idea that, at intermediate kinetic rate, the structure comprises (at percolation) some clusters that are more open/elongated clusters than at lower or higher kinetic rate. It is worth pointing out that, although the dimensionality value D becomes larger at higher or lower kinetic rate, it remains smaller than 3. This suggests that, even at low kinetic rate, the clusters remain partially loosely packed and elongated—in accordance with the clusters are not fully spherical (i.e., with a ~0.5 average aspect ratio, see Fig. 2-8(c)). Overall,

these results highlight how the mesoscale topology of C–S–H gels is closely correlated to the kinetics of its precipitation.



(a)



(b)

Figure 2-9: (a) Logarithm of the cluster size N_g (i.e., expressed in terms of the number of grains in cluster) at the percolation threshold (i.e., $\varphi = \varphi_c$) as a function of the logarithm of the cluster radius of gyration R_g for select kinetic rates R. The lines are power law fits (Eq. (2-6)). **(b)** Fractal dimension D (i.e., the slope of the lines in panel (a), see Eq. (2-6)) as a function of the kinetic rate R.

2.5 Conclusions

To summarize, this study constitutes an attempt to decipher the role of structural percolation in controlling the precipitation kinetics and thermodynamics of C–S–H gels. To this end, we conduct a series of coarse-grained grand canonical Monte Carlo simulations modeling the formation of a C–S–H gel with varying precipitation rates. We demonstrate the existence of a close correlation between structural percolation and both the kinetics and thermodynamics of C–S–H's precipitation. On the one hand, we find that the accelerating-to-decelerating transition occurs when the gel structure exhibit percolation. On the other hand, the precipitation kinetics is found to strongly affect the mesostructure of the gel at the percolation threshold. As such, these results

highlight the complex, retroactive relationship between the structure and kinetics of gels. More generally, these results suggest that the mesostructure of gels can be finely tuned by carefully adjusting the precipitation kinetics, and vice versa. This paves the way toward the rational design of tailored gels that can set “on-demand” with tunable degrees of packing.

2.6 References

[1] M.Y. Lin, H.M. Lindsay, D.A. Weitz, R.C. Ball, R. Klein, P. Meakin, Universality in colloid aggregation, Nature. 339 (1989) 360–362. https://doi.org/10.1038/339360a0.

[2] V. Trappe, V. Prasad, L. Cipelletti, P.N. Segre, D.A. Weitz, Jamming phase diagram for attractive particles, Nature. 411 (2001) 772–775. https://doi.org/10.1038/35081021.

[3] H. Liu, T. Du, N.M.A. Krishnan, H. Li, M. Bauchy, Topological optimization of cementitious binders: Advances and challenges, Cement and Concrete Composites. (2018). https://doi.org/10.1016/j.cemconcomp.2018.08.002.

[4] K. Ioannidou, R. J.-M. Pellenq, E.D. Gado, Controlling local packing and growth in calcium–silicate–hydrate gels, Soft Matter. 10 (2014) 1121–1133. https://doi.org/10.1039/C3SM52232F.

[5] K. Ioannidou, K.J. Krakowiak, M. Bauchy, C.G. Hoover, E. Masoero, S. Yip, F.-J. Ulm, P. Levitz, R.J.-M. Pellenq, E.D. Gado, Mesoscale texture of cement hydrates, PNAS. 113 (2016) 2029–2034. https://doi.org/10.1073/pnas.1520487113.

[6] E. Masoero, E. Del Gado, R.J.-M. Pellenq, F.-J. Ulm, S. Yip, Nanostructure and Nanomechanics of Cement: Polydisperse Colloidal Packing, Phys. Rev. Lett. 109 (2012) 155503. https://doi.org/10.1103/PhysRevLett.109.155503.

[7] G. Constantinides, F.-J. Ulm, The nanogranular nature of C–S–H, Journal of the Mechanics and Physics of Solids. 55 (2007) 64–90. https://doi.org/10.1016/j.jmps.2006.06.003.

[8] M. Vandamme, F.-J. Ulm, P. Fonollosa, Nanogranular packing of C–S–H at substochiometric conditions, Cement and Concrete Research. 40 (2010) 14–26. https://doi.org/10.1016/j.cemconres.2009.09.017.

[9] E. Masoero, E.D. Gado, R. J.-M. Pellenq, S. Yip, F.-J. Ulm, Nano-scale mechanics of colloidal C–S–H gels, Soft Matter. 10 (2014) 491–499. https://doi.org/10.1039/C3SM51815A.

[10] C. Le Quéré, R.J. Andres, T. Boden, T. Conway, R.A. Houghton, J.I. House, G. Marland, G.P. Peters, G. van der Werf, A. Ahlström, R.M. Andrew, L. Bopp, J.G. Canadell, P. Ciais, S.C. Doney, C. Enright, P. Friedlingstein, C. Huntingford, A.K. Jain, C. Jourdain, E. Kato, R.F. Keeling, K. Klein Goldewijk, S. Levis, P. Levy, M. Lomas, B. Poulter, M.R. Raupach, J. Schwinger, S. Sitch, B.D. Stocker, N. Viovy, S. Zaehle, N. Zeng, The global carbon budget 1959–2011, Earth System Science Data Discussions. 5 (2012) 1107–1157. https://doi.org/info:doi:10.5194/essdd-5-1107-2012.

[11] M. Bauchy, Nanoengineering of concrete via topological constraint theory, MRS Bulletin. 42 (2017) 50–54. https://doi.org/10.1557/mrs.2016.295.

[12] K. Ioannidou, M. Kanduč, L. Li, D. Frenkel, J. Dobnikar, E. Del Gado, The crucial effect of early-stage gelation on the mechanical properties of cement hydrates, Nature Communications. 7 (2016) 12106. https://doi.org/10.1038/ncomms12106.

[13] H. Liu, S. Dong, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Effects of polydispersity and disorder on the mechanical properties of hydrated silicate gels, Journal of the Mechanics and Physics of Solids. 122 (2019) 555–565. https://doi.org/10.1016/j.jmps.2018.10.003.

[14] H. Liu, S. Dong, N.M.A. Krishnan, E. Masoero, G. Sant, M. Bauchy, Long-term creep deformations in colloidal calcium–silicate–hydrate gels by accelerated aging simulations, Journal of Colloid and Interface Science. 542 (2019) 339–346. https://doi.org/10.1016/j.jcis.2019.02.022.

[15] H. Manzano, E. Masoero, I. Lopez-Arbeloa, H. M. Jennings, Shear deformations in calcium silicate hydrates, Soft Matter. 9 (2013) 7333–7341. https://doi.org/10.1039/C3SM50442E.

[16] H. Manzano, S. Moeini, F. Marinelli, A.C.T. van Duin, F.-J. Ulm, R.J.-M. Pellenq, Confined Water Dissociation in Microporous Defective Silicates: Mechanism, Dipole Distribution, and Impact on Substrate Properties, J. Am. Chem. Soc. 134 (2012) 2208–2215. https://doi.org/10.1021/ja209152n.

[17] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[18] D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, Elsevier, 2001.

[19] M. Corsten, N. Jan, R. Jerrard, Critical properties of random-site percolation in two and three dimensions: A Monte Carlo study, Physica A: Statistical Mechanics and Its Applications. 156 (1989) 781–794. https://doi.org/10.1016/0378-4371(89)90020-4.

[20] N. Tsakiris, M. Maragakis, K. Kosmidis, P. Argyrakis, Percolation of randomly distributed growing clusters: Finite-size scaling and critical exponents for the square lattice, Physical Review E. 82 (2010). https://doi.org/10.1103/PhysRevE.82.041108.

[21] G.D. Scott, D.M. Kilgour, The density of random close packing of spheres, J. Phys. D: Appl. Phys. 2 (1969) 863. https://doi.org/10.1088/0022-3727/2/6/311.

[22] A. Donev, I. Cisse, D. Sachs, E.A. Variano, F.H. Stillinger, R. Connelly, S. Torquato, P.M. Chaikin, Improving the Density of Jammed Disordered Packings Using Ellipsoids, Science. 303 (2004) 990–993. https://doi.org/10.1126/science.1093010.

[23] K. Vollmayr, W. Kob, K. Binder, Cooling-rate effects in amorphous silica: A computer-simulation study, Physical Review B. 54 (1996) 15808–15827. https://doi.org/10.1103/PhysRevB.54.15808.

[24] P.G. Debenedetti, F.H. Stillinger, Supercooled liquids and the glass transition, Nature. (2001). https://doi.org/10.1038/35065704.

[25] M. Abuhaikal, K. Ioannidou, T. Petersen, R.J.-M. Pellenq, F.-J. Ulm, Le Châtelier's conjecture: Measurement of colloidal eigenstresses in chemically reactive materials, Journal of the Mechanics and Physics of Solids. 112 (2018) 334–344. https://doi.org/10.1016/j.jmps.2017.12.012.

[26] A.P. Thompson, S.J. Plimpton, W. Mattson, General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions, The Journal of Chemical Physics. 131 (2009) 154107. https://doi.org/10.1063/1.3245303.

[27] N.M.A. Krishnan, B. Wang, Y. Yu, Y. Le Pape, G. Sant, M. Bauchy, Enthalpy Landscape Dictates the Irradiation-Induced Disordering of Quartz, Physical Review X. 7 (2017). https://doi.org/10.1103/PhysRevX.7.031019.

[28] M. Bauchy, M. Wang, Y. Yu, B. Wang, N.M.A. Krishnan, E. Masoero, F.-J. Ulm, R. Pellenq, Topological Control on the Structural Relaxation of Atomic Networks under Stress, Physical Review Letters. 119 (2017). https://doi.org/10.1103/PhysRevLett.119.035502.

[29] P. Boolchand, D.G. Georgiev, B. Goodman, Discovery of the Intermediate Phase in Chalcogenide Glasses, Journal of Optoelectronics and Advanced Materials. 3 (2001) 703–720.

[30] P. Boolchand, B. Goodman, Glassy materials with enhanced thermal stability, MRS Bulletin. 42 (2017) 23–28. https://doi.org/10.1557/mrs.2016.300.

[31] P. Boolchand, M. Bauchy, M. Micoulaut, C. Yildirim, Topological Phases of Chalcogenide Glasses Encoded in the Melt Dynamics, Physica Status Solidi (b). 255 (2018) 1800027. https://doi.org/10.1002/pssb.201800027.

[32] M. Micoulaut, Simple Clues and Rules for Self-Organized Rigidity in Glasses, Journal of Optoelectronics and Advanced Materials. 9 (2007) 3235–3240.

[33] M.V. Chubynsky, M.-A. Brière, N. Mousseau, Self-organization with equilibration: A model for the intermediate phase in rigidity percolation, Physical Review E. 74 (2006). https://doi.org/10.1103/PhysRevE.74.016116.

[34] V. Morin, F. Cohen-Tenoudji, A. Feylessoufi, P. Richard, Evolution of the capillary network in a reactive powder concrete during hydration process, Cement and Concrete Research. 32 (2002) 1907–1914. https://doi.org/10.1016/S0008-8846(02)00893-1.

[35] M. Fourmentin, P. Faure, S. Gauffinet, U. Peter, D. Lesueur, D. Daviller, G. Ovarlez, P. Coussot, Porous structure and mechanical strength of cement-lime pastes during setting, Cement and Concrete Research. 77 (2015) 1–8. https://doi.org/10.1016/j.cemconres.2015.06.009.

Section A. Physics-Driven Computational Simulations: Make the Physics Simple

A2. Coarse-Grained Molecular Dynamics Simulation

Chapter 3. Effects of Polydispersity and Disorder on the Mechanical Properties of Hydrated Silicate Gels

3.1 Introduction

Colloidal gels—i.e., systems made of interacting grains with sizes ranging from nanometers to hundreds of nanometers [1,2]—are widely used in many realms of science and technology [3–5]. Colloid aggregation and gelation is controlled by a complex interplay between thermodynamics and kinetics [6–8]. After gelation, the structure of colloidal gels largely depends on the formation process [7,9]. In particular, the degree of polydispersity (i.e., the distribution of grain sizes) and the level of disorder affect the mechanical properties of colloidal gels—although such linkages remain poorly understood [3,10,11].

Calcium–silicate–hydrate (C–S–H) gel—the glue of concrete that forms upon the hydration of cement—is a typical inorganic colloidal hydrogel material [3,12,13]. The C–S–H phase largely controls the strength of cement paste and concrete [14,15]. This is significant as, due to its large carbon impact [16–18], there is a strong interest in improving concrete's mechanical properties—i.e., so that less material can be used while achieving constant performance. In turn, the colloidal structure of C–S–H largely controls its mechanical response [10,19,20]. At the mesoscale, C–S–H forms a polydisperse, disordered gel-like structure, with the grain size ranging from nanometers to several tens nanometers [3,12,13]. However, the linkages between the structure and mechanical behavior of C–S–H are not fully elucidated.

Recently, Masoero *et al.* introduced a polydisperse colloidal model of C–S–H that shows a good agreement with nanoindentation experiments [3]. Based on this model, the packing density was shown to have a first order effect on the mechanical properties of C–S–H [3,10]. In the same spirit, Ioannidou e*t al.* proposed an alternative model to account for the early-age structure and

properties of C–S–H [14]. These colloidal models have been widely investigated to elucidate the relationship between C–S–H's structure and mechanical properties [10,13,21,22]. However, several questions, in this regard, remain unanswered. Which structural features have a first-order effect and which one do not? What is the effect of polydispersity and structural disorder on the mechanical properties of C–S–H? Indeed, although polydispersity has been shown to play a critical role in controlling the packing density (and, thereby, the mechanical properties) [3,10], the role of polydispersity at constant packing density remains unclear. Similarly, although structural and mechanical heterogeneity has been pointed out to impact the nanomechanics of C–S–H gels [10,22,23], the role of the extent of order and disorder on macroscopic properties remains poorly understood. Meanwhile, little is known about the role of the level of order in the mesostructure of C–S–H—which has been suggested to be higher in high-density C–S–H phases [14,19,22,24].

To address these questions, we conduct some grand canonical Monte Carlo (GCMC) simulations to investigate the effect of grain polydispersity and structural disorder on the mechanical properties of colloidal C–S–H, by relying on the model of Masoero *et al* [3]. Our simulations yield an excellent agreement with nanoindentation data for a wide range of packing density. We show that, at constant packing fraction, polydispersity does not affect the stiffness and hardness of C–S–H. In contrast, we demonstrate that the degree of disorder has a first-order effect on the mechanical properties of C–S–H gels. This is ascribed to the existence of some local stress heterogeneity within the disordered structure, which arises from the out-of-equilibrium nature of the C–S–H gels. We show that, upon loading, such stress heterogeneity induces the occurrence of some local structural nanoyielding, which, in turn, decreases the apparent macroscopic stiffness of the bulk C–S–H gel. These results highlight the critical importance of the level of order and disorder in the mechanical properties of colloidal systems.

This paper is organized as follows. In Sec. 3.2, we describe the simulation methods used to generate the C–S–H gel configurations and to calculate their mechanical properties. In Sec. 3.3, we compare the outcomes of our simulations to nanoindentation data and investigate the effect of polydispersity and disorder on C–S–H's mechanical properties. These results are discussed in Sec. 3.4. Finally, some conclusions are given in Sec. 3.5.

3.2 Methods

3.2.1 Preparation of the C–S–H configurations

To establish our conclusions, we adopt here the colloidal model of C–S–H introduced by Masoero *et al* [3,10]. In this model, the C–S–H gel is described as an ensemble of polydisperse spherical grains that interact with each other via a generalized Lennard-Jones interaction energy potential:

$$U_{ij}(r_{ij}) = 4\varepsilon(\sigma_i, \sigma_j)\left[\left(\frac{\bar{\sigma}_{ij}}{r_{ij}}\right)^{2\alpha} - \left(\frac{\bar{\sigma}_{ij}}{r_{ij}}\right)^{\alpha}\right] \qquad \textbf{Eq. (3-1)}$$

where σ_i and σ_j are the diameters of grains i and j, $\bar{\sigma}_{ij} = (\sigma_i + \sigma_j)/2$ is the average diameter for a given pair of atom, α is a parameter that controls the narrowness of the potential well, r_{ij} is distance between the centers of the grains i and j, and $\varepsilon(\sigma_i, \sigma_j)$ is depth of the potential energy well. By considering each pair of grains in contact as two springs in series, the depth is given by $\varepsilon(\sigma_i, \sigma_j) = A_0 \beta_{ij} {\bar{\sigma}_{ij}}^3$, where $A_0 = kE$ is a prefactor that is proportional to the bulk Young's modulus E of a grain, wherein $k = 0.002324$ (computed by the serial spring model) and $E = 63.6$ GPa (based on previous atomistic simulations of bulk C–S–H) [10,25]. $\beta_{ij} = \sigma_i \sigma_j / {\bar{\sigma}_{ij}}^2$ is a correction term arising from the serial arrangement. The potential defined in Eq. (3-1) shows a minimum at $r_m = \sqrt[\alpha]{2}\bar{\sigma}_{ij}$ so that the effective diameter of a grain i is defined as $\sigma_{0,i} = \sqrt[\alpha]{2}\sigma_i$. The

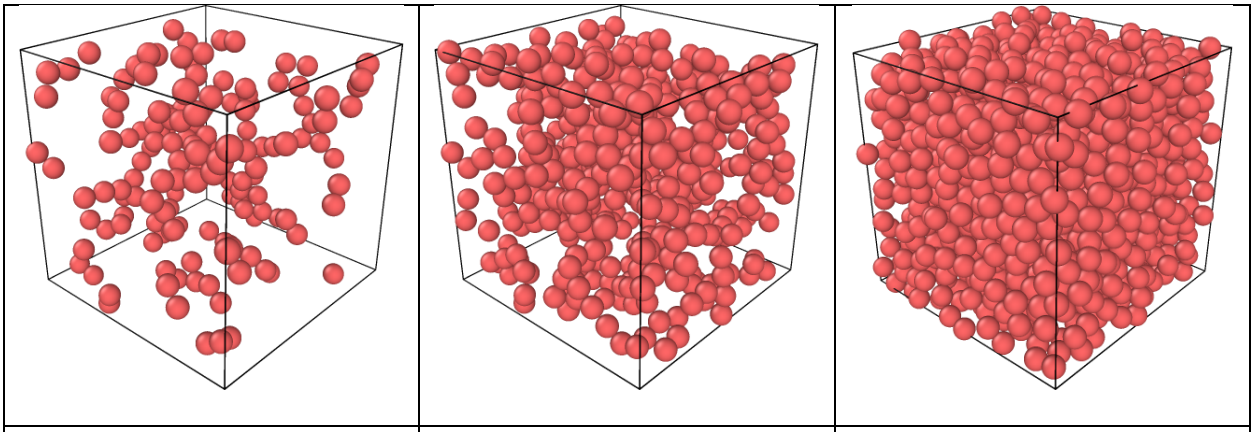
attractive force is maximum at a distance $r_u = \sqrt[\alpha]{\frac{4\alpha+2}{\alpha+1}}\bar{\sigma}_{ij}$ so that, by choosing $\alpha = 14$, the tensile strain at failure $\varepsilon_u = (r_u - r_m)/r_m$ is close to the value of 5% obtained in previous atomistic simulation of bulk C–S–H [10,25,26]. Note that orientational effects arising from the layered structure of bulk C–S–H are not accounted by present the two-body potential—we assume here that, overall, the grain anisotropy is lost due to the random mutual orientation of the C–S–H grains [10,27,28].

The C–S–H configurations are generated by grand canonical Monte Carlo (GCMC) simulations, as described in the following. Starting from an initially empty cubic box of size 1000 Å, some C–S–H grains are iteratively inserted, wherein the size of each grain is randomly selected from a uniform distribution between a minimum σ_m and a maximum σ_M value. The standard deviation μ of the distribution is then used to define the polydispersity index of the configuration as:[3,10]

$$\delta = \mu/[(\sigma_m + \sigma_M)/2] \qquad \textbf{Eq. (3-2)}$$

Here, various polydispersity values are considered, with σ ranging from 3.0 to 35 nm, and the number of grains at saturation ranging from 4000 to 12000. In detail, each GCMC step comprises *X* attempts of grain insertions or deletions followed by *Y* attempts to randomly displace an existing grain. At each step, the probability of success of the attempt is given by $\exp(-\Delta U/k_{\mathrm{B}}T)$, where k_{B} is the Boltzmann constant, *T* the temperature, and ΔU is the variation in potential energy caused by the trial insertion/displacement [3,10,13]. The factor $R = X/Y$ is then qualitatively equivalent to a precipitation rate, which characterizes the time duration during which the grains are allowed to reorganize in between two successive insertions. Namely, a large R value corresponds to a high precipitation rate, wherein the grains have only limited opportunity to move during precipitation. In this work, we use $R = 0.01$. This process is iteratively repeated until the number of inserted

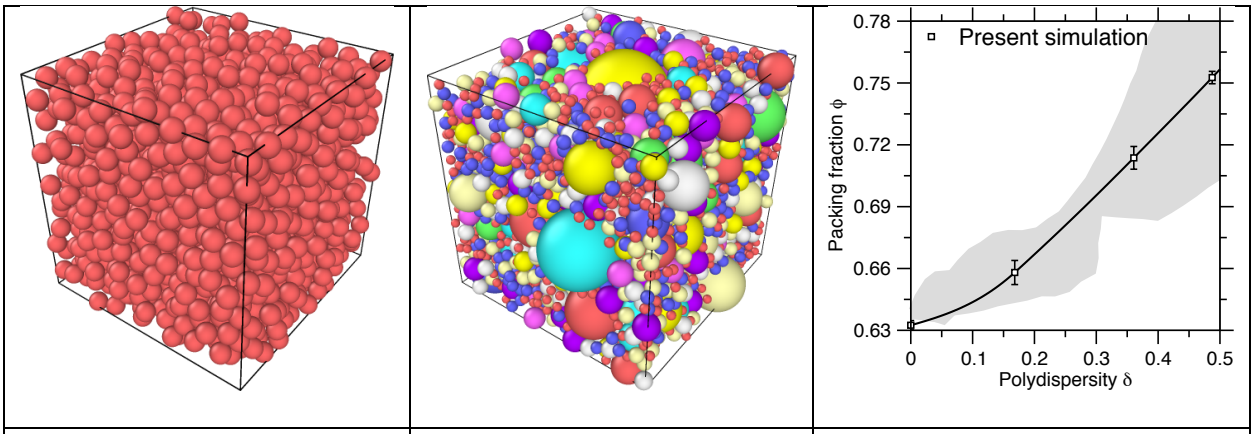
grains reaches a plateau (see Fig. 3-1). The packing fraction ϕ of each configuration is computed as $\phi = \sum_i \pi/6{\sigma_{0,i}}^3/V$, where V is the volume of the simulation box.



(a) **(b)** **(c)**

Figure 3-1: Snapshots of select monodisperse C–S–H gel configurations with increasing packing fraction values of **(a)** 0.055, **(b)** 0.174 and **(c)** 0.417.

In agreement with previous simulations [3,29], we observe that the C–S–H models exhibiting higher degree of polydispersity eventually reach higher final packing fraction values—as small grains are able to fill the space left in between larger grains (see Fig. 3-2). For monodisperse configurations, the packing fraction at saturation is around 0.63, that is, close to the theoretical packing limit of random monodisperse spheres [30,31]. Further, we note that the evolution of the packing fraction at saturation with polydispersity is in good agreement with the range of data previously observed [3,29]. Note that this analysis is conducted on five independent simulations of precipitation performed for each polydispersity to calculate the mean value and standard deviation of the final packing fraction.



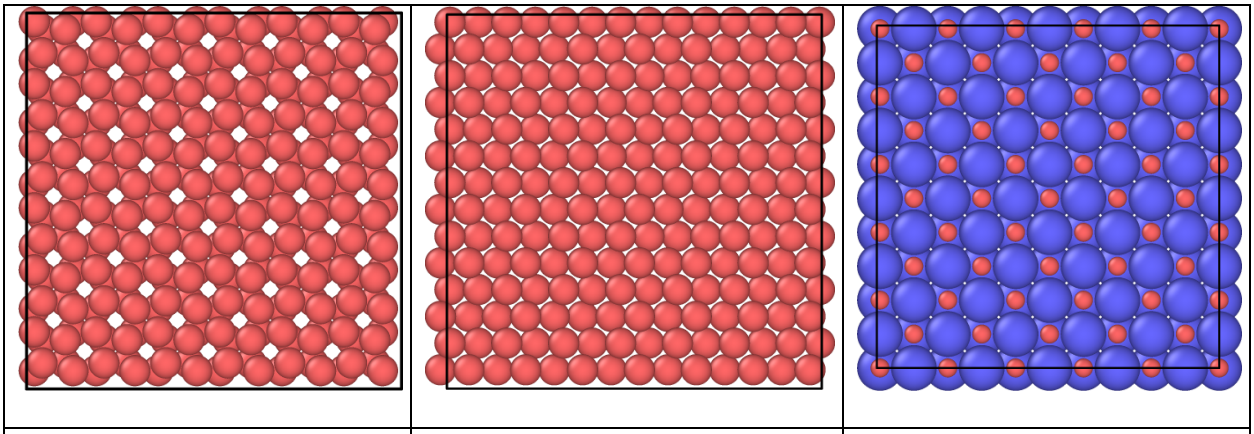
(a) **(b)** **(c)**

Figure 3-2: Snapshots of select **(a)** monodisperse (polydispersity index $\delta = 0$) and **(b)** polydisperse C–S–H gel configurations ($\delta = 0.49$). **(c)** Computed packing fraction (at saturation) of the C–S–H gel configurations as a function of the polydispersity index (see Eq. (3-2)). The grey area indicates the range of data previously observed [3,29].

3.2.2 Preparation of artificial C–S–H crystalline configurations

To assess the effect of order and disorder on the mechanical properties of C–S–H, a selection of "artificial" C–S–H crystalline configurations with varying different packing fractions are generated. This is achieved by creating a series of C–S–H configurations based on a selection of crystalline lattices and relaxing the configuration at zero stress to assess the stability of the crystal. Note that the same interatomic potential (Eq. (3-1)) is used for the ordered C–S–H configurations. A series of five stable artificial C–S–H crystals is considered herein, namely, (1) a DNA-like structure ($\phi = 0.63$, $\delta = 0$), (2) a CsCl-type structure ($\phi = 0.73$, $\delta = 0.15$), (3) a body-centered cubic structure (HCP) structure ($\phi = 0.74$, $\delta = 0$), (4) a face-centered cubic structure (FCC) structure ($\phi = 0.74$, $\delta = 0$), (5) and a NaCl-type structure ($\phi = 0.79$, $\delta = 0.41$)—see Fig. 3-3. Note that the crystals 1, 3, and 4 are monodisperse, with a grain size of 5 nm. Here, the DNA-like

structure refers to a type of helical lattice structure where grains are aligned on the lattices—a configuration that is obtained by removing select grains from a HCP configuration (see Fig. 3-3a). Note that the simple cubic lattice is found to be an unstable configuration for C–S–H. In contrast to the previous configurations, the configurations 2 and 5 mimic oxide crystals, i.e., wherein small cations fill the available space left in between the O atoms. Based on the same idea, the CsCl- and NaCl-like configurations are obtained by introducing smaller grains (3.66 and 3.31 nm, respectively) in between larger grains (5 and 8 nm, respectively). This yields some polydisperse crystalline configurations exhibiting large packing fraction values.



(a)	**(b)**	**(c)**

Figure 3-3: Snapshots of some of the artificial C–S–H crystalline configurations generated herein, which comprise (in order of increasing packing fraction ϕ) **(a)** a DNA-like structure (ϕ = 0.63), **(b)** a HCP crystal (ϕ = 0.74), and **(c)** a NaCl-type crystal (ϕ = 0.79).

3.2.3 Stiffness computation

Select C–S–H configurations corresponding to different packing fractions are extracted during the GCMC process. These structures are relaxed by molecular dynamics simulations in the *NVT* ensemble, and subjected to an energy minimization prior to any subsequent characterization.

In addition, for each degree of polydispersity considered herein, final C–S–H configurations (i.e., at saturation) are relaxed by molecular dynamics simulations in the *NPT* ensemble at zero stress and subjected to a final energy minimization. The stiffness tensor of each configuration is then computed by subjecting the simulation box to a series of axial and shear plane deformations along each Cartesian axis [32,33]. The maximum strain for each deformation is restricted to ± 0.005. The corresponding changes in the potential energy ∂U and strain ∂e define six stress components:

$$s_\alpha = \frac{1}{V}\frac{\partial U}{\partial e_\alpha} \quad \textbf{Eq. (3-3)}$$

and 36 elastic constants:

$$C_{\alpha\beta} = \frac{1}{V}\frac{\partial^2 U}{\partial e_\alpha e_\beta} \quad \textbf{Eq. (3-4)}$$

where α and β are the Cartesian direction indexes. All configurations are found to be nearly fully isotropic. The Young's modulus (E), shear modulus (G), bulk modulus (K), and Poisson's ratio (ν) are then calculated from the stiffness tensor. Finally, the indentation modulus (M) is determined as [19,34]:

$$M = 4G\frac{3K+G}{3K+4G} \quad \textbf{Eq. (3-5)}$$

3.2.4 Hardness computation

The indentation hardness (H) of the relaxed C–S–H configurations is then computed following the method introduced by Qomi *et al.* [35–37], as described in the following. This method is based on the computation of the failure envelope of the system as a function of the normal and shear stresses (σ, $\boldsymbol{\tau}$). In detail, the failure envelope is determined by performing a series of different deformations (pure uniaxial tension and compression, pure shear, and combinations thereof) by incrementally increasing the strain via a series of box deformation. For each

deformation, the yield stress is determined based on the 0.2% offset method. In each case, the normal and shear stresses at the yield point are used to draw a Mohr circle. The envelope of all the Mohr circles is then fitted by a Mohr-Coulomb failure criterion:

$$\tau = C - \sigma \times \tan\varphi \qquad \textbf{Eq. (3-6)}$$

where C is the cohesion stress φ is the friction angle. The hardness H of a cohesive-frictional material such as C–S–H is then given by:

$$\frac{H}{C} = \frac{\delta(\varphi,\theta)}{\tan\varphi} = \frac{1}{\tan\varphi}\sum_{k=1}^{N}[a_k(\theta)\tan\varphi]^k \qquad \textbf{Eq. (3-7)}$$

where θ is the indenter apex angle, a_k are fitting parameters for a given indenter geometry, and N is the maximum order of the polynomial expansion. In the case of the C–S–H gels, since the friction angle is on the order of 5° or less, $\tan\varphi$ is here considered negligible, and, as a result, the hardness is approximated as $5.8C$ [35]. More details about the hardness computation methodology can be found in Ref [35].

3.2.5 Stress per grain

Finally, the local stress experienced by each C–S–H grain is computed. Although stress is intrinsically a macroscopic property that is ill-defined for individual grains, a local "stress per grain" can be defined based on the formalism proposed by Thompson *et al* [38]. This approach consists in expressing the contribution of each grain *i* to the virial of the system [39]:

$$3\sigma_i V_i = m_i v_i^2 + \vec{r_i}.\vec{F_i} \qquad \textbf{Eq. (3-8)}$$

where σ_i is the local stress per grain, V_i, m_i, v_i, and $\vec{r_i}$ are the volume, mass, velocity, and position of the grain *i*, respectively, and $\vec{F_i}$ is the resultant of the force applied on the grain *i* by all the other grains in the system. Here, we define the volume V_i of each grain based on its Voronoi volume. By convention, a positive stress represents here a state of tension, whereas a negative one

represents a state of compression. We recently used this approach to quantify the internal stress exhibited by stressed–rigid atomic networks [40] and mixed alkali glasses [39,41–43]. It should be noted that, in the thermodynamic sense, stress is only properly defined for a large ensemble of atoms, so the physical meaning of the "stress per grain" is unclear. Nevertheless, this quantity can conveniently capture the existence of local instabilities within the gel due to competitive inter-grain forces [23].

3.3 Results

3.3.1 Comparison with nanoindentation experiments

We first compare the computed modulus and hardness values with nanoindentation experimental data [20,44]. Overall, as shown in Fig. 3-4, we obtain an excellent agreement between simulation and nanoindentation data over a large range of packing fractions, which establishes the ability of our simulations to properly describe the nanomechanics of C–S–H. We observe that both the indentation modulus and hardness increase monotonically with the packing fraction. In details, the increase is noted to be limited at low packing fraction ($\phi < 0.5$), whereas it is more pronounced at high packing fraction ($\phi > 0.5$)—in agreement with previous simulation data [13]. The specific value of $\phi = 0.5$ corresponds to the percolation point of the system, which supports the idea of a granular description of C–S–H, wherein the grains interact with each other via Hertz contact points [19,20]. However, we note that a finite value of the indentation modulus is achieved even below the percolation point. This highlights the colloidal nature of C–S–H and the fact that the grains exhibit some level of attraction even when they are not in direct contact with each other.

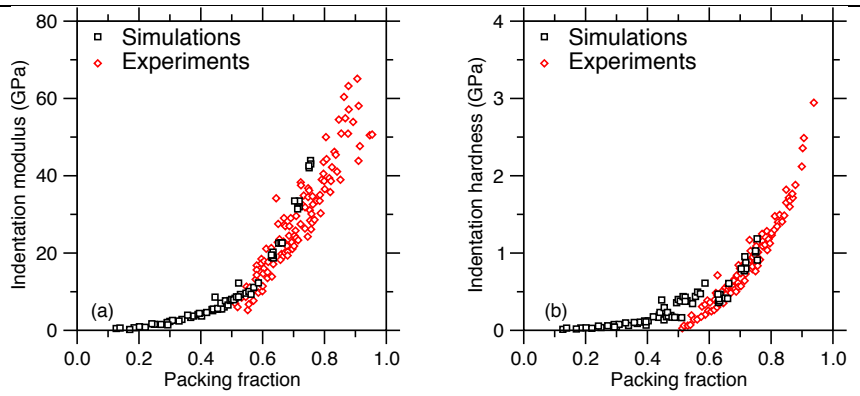


Figure 3-4: Computed **(a)** indentation modulus and **(b)** hardness of C–S–H gels as a function of the packing fraction. The results are compared with experimental nanoindentation data [20,44].

3.3.2 Effect of polydispersity

We now investigate the effect of the polydispersity in the grain sizes on the nanomechanics of C–S–H gels. To this end, Fig. 3-5 shows the computed values of the indentation modulus and hardness as a function of the packing fraction for select polydispersity (see Eq. (3-2)). Overall, we observe that larger polydispersity values yield larger final packing fraction at saturation (see Fig. 3-2c) and, thereby, larger values of indentation modulus and hardness. However, we note that, at constant packing fraction, the mechanical properties of C–S–H do not depend on polydispersity since all data fall on the same "master curve" (see Fig. 3-5). This suggests that, in the case of the present disordered C–S–H configurations, the packing fraction is the only order parameter that controls the stiffness and hardness of C–S–H, whereas polydispersity itself is not a relevant parameter. This result is in agreement with previous observations reporting some universal scaling relationships between density and stiffness [45–50].

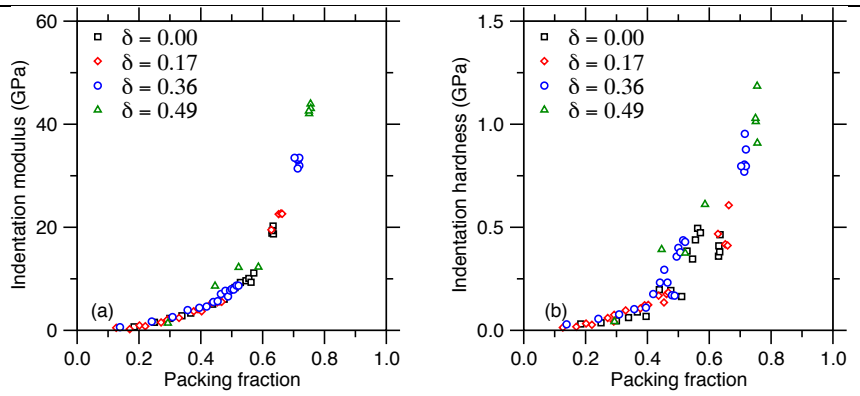


Figure 3-5: Computed **(a)** indentation modulus and **(b)** hardness of C–S–H gels with varying polydispersity index (δ, see Eq. (3-2)) as a function of packing fraction.

3.3.2 Effect of disorder

Finally, we focus on elucidating the effect of disorder on the nanomechanics of C–S–H. To this end, Fig. 3-6 shows the indentation modulus and hardness of the disordered C–S–H configurations (i.e., obtained by GCMC simulations) and those of the artificial C–S–H crystals. First, we observe a significant shift between the indentation modulus and hardness of the disordered and ordered C–S–H configurations, wherein ordered configurations systematically exhibit higher values than their disordered counterparts at constant packing fraction. Second, we note that the indentation modulus and hardness of the crystalline C–S–H configurations significantly departs from the experimental data, which supports the intrinsically disordered nature of C–S–H gels. Last, we note that, in contrast to the case of the disordered C–S–H configuration, the packing fraction is not a universal order parameter for the ordered configurations. Indeed, for instance, we note that, even though the FCC and HCP configurations have the same packing fraction (ϕ = 0.74), the HCP configuration shows significantly larger indentation modulus and hardness values than those of the FCC configuration. This demonstrates that the details of the structure have a strong impact on the mechanical properties of ordered configurations. Overall,

these results highlight the critical role played by the degree of order and disorder in controlling the mechanical properties of C–S–H gels.

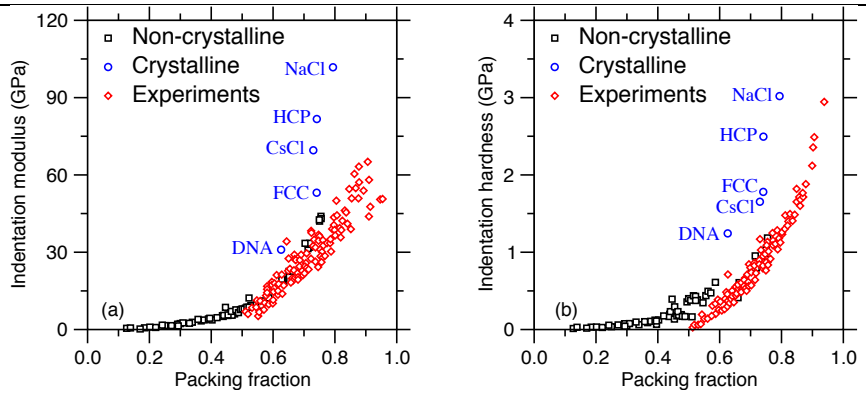


Figure 3-6: Computed **(a)** indentation modulus and **(b)** hardness of disordered (back squares) and crystalline (blue circles) C–S–H gels as a function of the packing fraction. Experimental nanoindentation data (red diamonds) are added for comparison [20,44].

3.4 Discussion

3.4.1 Effect of disorder on stiffness

We now discuss the origin of the results presented in Sec. 3.3 and further investigate the effect of order and disorder on the stiffness of the C–S–H gels. Fig. 3-7 shows the Young's modulus, shear modulus, bulk modulus, and Poisson's ratio of the ordered and disordered C–S–H configurations as a function of the packing fraction. First, we note that all the moduli increase monotonically with the packing fraction (see Figs. 3-7a, 3-7b, and 3-7c), in agreement with the fact that lower porosity results in higher stiffness [45]. Further, we note that, similar to the indentation modulus, ordered C–S–H gels systematically exhibit higher Young's modulus and shear modulus values than their disordered counterparts at constant packing fraction (see Figs. 3-7a and 3-7b). Again, this suggests that an increased level of order tends to enhance stiffness.

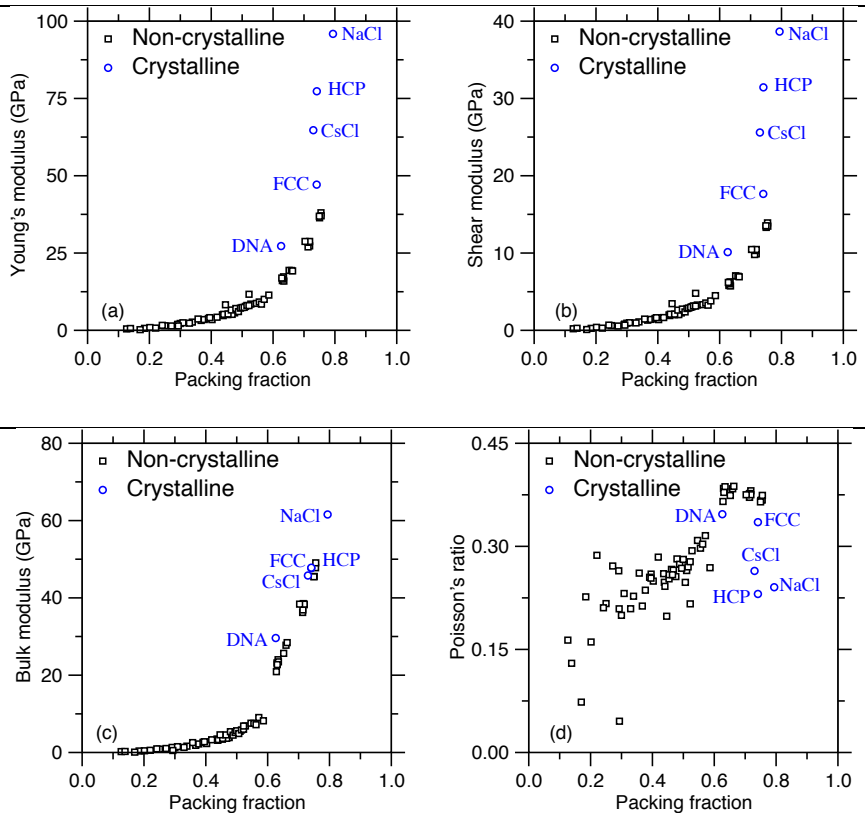


Figure 3-7: Computed **(a)** Young's modulus, **(b)** shear modulus, **(c)** bulk modulus, and **(d)** Poisson's ratio of the disordered (back squares) and crystalline (blue circles) C–S–H gels as a function of the packing fraction.

Second, we note that the Poisson's ratio also monotonically increases with increasing packing fractions (see Fig. 3-7d). Such a trend has been observed for a large variety of materials and has been suggested to arise from the fact that systems featuring higher packing fraction values have less ability to locally densify upon longitudinal loading and, hence, show a higher propensity for lateral deformations [45]. However, we observe a bifurcation between the Poisson's ratios of the ordered and disordered configurations, wherein ordered gels systematically exhibit a lower Poisson's ratio than their disordered counterparts at constant packing fraction. This suggests that

disordered systems exhibit a pronounced propensity for lateral deformations when subjected to longitudinal loads, which may arise from an increased mobility of the atoms.

In contrast, interestingly, we note that the bulk modulus values of the ordered and disordered configurations fall on the master curve (see Fig. 3-7c). This suggests that, in contrast to the other moduli, the level of disordered and the details of the structure do not impact the bulk modulus of gels—and that the packing density acts as an order parameter. This can be understood from the fact that the bulk modulus is a quantity that is primarily related to the volumic density of the potential energy between grains [51], which is similar for ordered and disordered configurations. Overall, this suggests that the level of disorder does not significantly affect the response of C–S–H when subjected to hydrostatic loads, but has a more pronounced effect in controlling its response to shear.

3.4.2 Nanoyielding and stress heterogeneity

Finally, we investigate the origin of the distinct behaviors of ordered and disordered C–S–H gels under shear. To the end, we focus on two select C–S–H gel configurations: (i) a disordered monodisperse C–S–H at saturation and (ii) a crystalline DNA-like C–S–H structure. These configurations are selected as they exhibit a different degree of order while having similar packing fraction values ($\phi = 0.63$). These two configurations are then subjected to pure shear deformation by gradually deforming the simulation box and performing an energy minimization after each incremental deformation. Figure 3-8 shows the corresponding stress-strain curves. First, we observe that, in both cases, the stress initially increases fairly linearly with strain until the system yields—which manifests by a plateau in the stress-strain curve (see Fig. 3-8). However, the slope of the stress-strain curve of the disordered C–S–H configuration is significantly lower than that of

the ordered one, in agreement with the fact that the disordered C–S–H structure exhibits a lower shear modulus (see Fig. 3-7b).

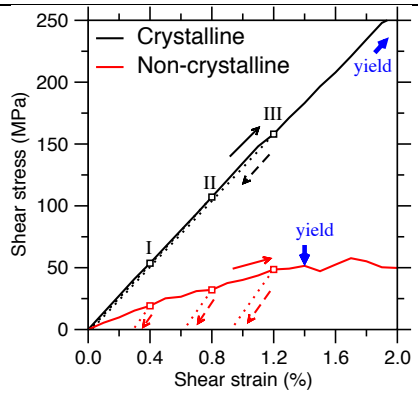


Figure 3-8: Computed shear stress vs. shear strain upon shearing in a disordered and crystalline C–S–H gel with similar packing fraction. In each case, three configurations (I, II, and III) achieved at select shear strains (0.4%, 0.8%, and 1.2%, respectively) are subsequently unloaded (dashed curves).

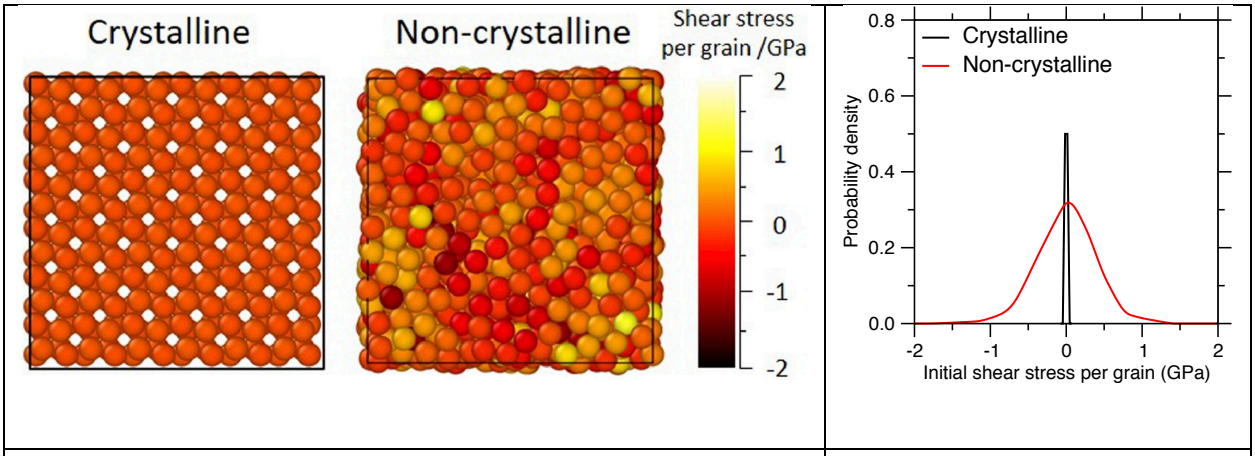
To assess the occurrence of any plastic deformations upon shearing, three configurations (I, II, and III in Fig. 3-8) obtained at select shear strains (0.4%, 0.8%, and 1.2%, respectively) are extracted and subsequently unloaded. We observe that, in the case of the ordered C–S–H system, the deformation is fully reversible, that is, the loading and unloading stress-strain curves are similar to each other—as expected in the case of a perfectly elastic deformation. In contrast, in the case of the disordered C–S–H gel, we observe a significant degree of irreversibility, that is, the unloading stress-strain curves differs from that obtained upon loading and the system remains permanently deformed after a loading/unloading cycle. Such irreversibility suggests that, although the whole system deforms in a linear fashion upon increasing stress, some irreversible plastic events upon loading occur in the structure. Note that, upon unloading, the stress-strain curve of the disordered C–S–H configurations exhibits a slope that is fairly similar to that of the ordered configuration.

This suggests that the plastic events that are activated upon loading are solely responsible for the difference in the shear modulus of the ordered and disordered configurations.

To better understand the nature of these plastic events, we compute the local shear stress per grain in both the ordered and disordered C–S–H configurations before any deformation is applied, that is, the overall structure being relaxed to zero stress. As expected and shown in Fig. 3-9, we observe that all grains are experiencing a nearly zero shear stress in the crystalline C–S–H configuration. In contrast, we observe the existence of some significant stress variations in the disordered C–S–H configuration, wherein the local shear stress per grain ranges from around –1 to +1 GPa. Note that this local stress does not result in any macroscopic stress, namely, the grains experiencing some positive or negative shear stress mutually compensate each other so that the overall structure remains at zero stress. The presence of such eigenstress is a manifestation of the out-of-equilibrium nature of disordered C–S–H and arises from the fact that the grain precipitate in a random fashion, so that the addition of new grains in a preexisting rigid structure necessarily involves some non-optimal contact among grains and, hence, the formation of some local stress [23]. A visual inspection of the local stress map (see Fig. 3-9a) reveals that the stress distribution is highly heterogeneous, that is, most of the stress is concentrated in some local clusters of interconnected grains.

Based on these observations, the following physical picture emerges. In crystalline C–S–H configurations, all the grains are initially at zero stress. Upon loading, the local stress experienced by each grain then increases in a homogeneous fashion. At a certain threshold stress, all the grains simultaneously reach their local yield stress, which causes the whole structure to yield. In contrast, in the disordered C–S–H configuration, the grains are initially pre-stressed (positively or negatively). The additional stress induced by the macroscopic deformation causes

some the positively pre-stressed grains to quickly reach their yield stress and, thereby, to release their local stress through local plastic, irreversible reorganizations. Note that such "nanoyielding" is highly localized (e.g., unlike collective plastic events like shear bands [52]) and, hence, does not result in the macroscopic yielding of the gel [10,21,53]. Overall, the nanoyielding events arising from the presence of some stress heterogeneity within the gel explain the origin of the lower apparent shear modulus of the disordered C–S–H configurations as well as their non-reversible behaviors.



(a) **(b)**

Figure 3-9: (a) Initial spatial distribution of the local shear stress per grain in a crystalline and disordered C–S–H gel with similar packing fraction before any deformation. **(b)** Distribution of the local shear stress per grain in the crystalline and disordered C–S–H gel configurations. The stress distribution of the disordered configuration is multiplied by 10× for readability.

3.4.3 Disordered nature of the structure of C–S–H gels

Finally, we briefly discuss the implications of the present findings. The C–S–H gels found in hydrated cement pastes have been categorized into various types, namely, low-density (LD) and high-density (HD) C–S–H. Originally, an underlying assumption of this categorization was that

low packing efficiency of LD C–S–H is due to its disordered nature, whereas the high packing efficiency of HD C–S–H arises from a crystal-like ordered structure [19,20,24]. Clearly, the fact that indentation modulus of the ordered C–S–H gels does not match with nanoindentation data (see Fig. 3-6) suggests that the variations in the packing density observed in the C–S–H gels forming in hydrated cement pastes (i.e., from LD to HD) does not arise from some variations in the level of order, but rather from different degree of polydispersity in grain sizes. These results also suggest that the yield stress of C–S–H gels (and, thereby, concrete strength) could be greatly enhanced by controlling the level of structural order and disorder and, hence, the extent of stress heterogeneity within the structure. The existence of local eigenstress within C–S–H is also likely to control the creep relaxation behavior of cement pastes [54,55].

3.5 Conclusions

Overall, these results reveal the crucial effect of structural disorder in controlling the mechanical properties of gels. This arises from the out-of-equilibrium nature of disordered gels, which results in the formation of some local size mismatch among neighboring grains and, thereby, the formation of some local eigenstress. In turn, the presence of such eigenstress result in the occurrence of the nanoyielding of individual grains upon loading, which reduces the effective stiffness and strength of the gel. In contrast, the mechanical properties are found to be independent of the extent of polydispersity in grain size. Overall, these results suggest that the mechanical properties of gels could be enhanced by tuning the extent of order and disorder in their structures.

3.6 References

[1] M.Y. Lin, H.M. Lindsay, D.A. Weitz, R.C. Ball, R. Klein, P. Meakin, Universality in colloid aggregation, Nature. 339 (1989) 360–362. https://doi.org/10.1038/339360a0.

[2] A.D. Dinsmore, D.A. Weitz, Direct imaging of three-dimensional structure and topology of colloidal gels, J. Phys.: Condens. Matter. 14 (2002) 7581. https://doi.org/10.1088/0953-8984/14/33/303.

[3] E. Masoero, E. Del Gado, R.J.-M. Pellenq, F.-J. Ulm, S. Yip, Nanostructure and Nanomechanics of Cement: Polydisperse Colloidal Packing, Phys. Rev. Lett. 109 (2012) 155503. https://doi.org/10.1103/PhysRevLett.109.155503.

[4] Wang Q., Wang L., Detamore M. S., Berkland C., Biodegradable Colloidal Gels as Moldable Tissue Engineering Scaffolds, Advanced Materials. 20 (2007) 236–239. https://doi.org/10.1002/adma.200702099.

[5] Y.M. Joshi, Dynamics of Colloidal Glasses and Gels, Annual Review of Chemical and Biomolecular Engineering. 5 (2014) 181–202. https://doi.org/10.1146/annurev-chembioeng-060713-040230.

[6] V.J. Anderson, H.N.W. Lekkerkerker, Insights into phase transition kinetics from colloid science, Nature. 416 (2002) 811–815. https://doi.org/10.1038/416811a.

[7] E. Zaccarelli, Colloidal gels: equilibrium and non-equilibrium routes, J. Phys.: Condens. Matter. 19 (2007) 323101. https://doi.org/10.1088/0953-8984/19/32/323101.

[8] P.J. Lu, E. Zaccarelli, F. Ciulla, A.B. Schofield, F. Sciortino, D.A. Weitz, Gelation of particles with short-range attraction, Nature. 453 (2008) 499–503. https://doi.org/10.1038/nature06931.

[9] V. Trappe, P. Sandkühler, Colloidal gels—low-density disordered solid-like states, Current Opinion in Colloid & Interface Science. 8 (2004) 494–500. https://doi.org/10.1016/j.cocis.2004.01.002.

[10] E. Masoero, E.D. Gado, R. J.-M. Pellenq, S. Yip, F.-J. Ulm, Nano-scale mechanics of colloidal C–S–H gels, Soft Matter. 10 (2014) 491–499. https://doi.org/10.1039/C3SM51815A.

[11] H. Tong, P. Tan, N. Xu, From Crystals to Disordered Crystals: A Hidden Order-Disorder Transition, Scientific Reports. 5 (2015) 15378. https://doi.org/10.1038/srep15378.

[12] H.M. Jennings, Refinements to colloid model of C-S-H in cement: CM-II, Cement and Concrete Research. 38 (2008) 275–289. https://doi.org/10.1016/j.cemconres.2007.10.006.

[13] K. Ioannidou, K.J. Krakowiak, M. Bauchy, C.G. Hoover, E. Masoero, S. Yip, F.-J. Ulm, P. Levitz, R.J.-M. Pellenq, E.D. Gado, Mesoscale texture of cement hydrates, PNAS. 113 (2016) 2029–2034. https://doi.org/10.1073/pnas.1520487113.

[14] K. Ioannidou, M. Kanduč, L. Li, D. Frenkel, J. Dobnikar, E. Del Gado, The crucial effect of early-stage gelation on the mechanical properties of cement hydrates, Nature Communications. 7 (2016) 12106. https://doi.org/10.1038/ncomms12106.

[15] J.W. Bullard, H.M. Jennings, R.A. Livingston, A. Nonat, G.W. Scherer, J.S. Schweitzer, K.L. Scrivener, J.J. Thomas, Mechanisms of cement hydration, Cement and Concrete Research. 41 (2011) 1208–1223. https://doi.org/10.1016/j.cemconres.2010.09.011.

[16] M. Bauchy, Nanoengineering of concrete via topological constraint theory, MRS Bulletin. 42 (2017) 50–54. https://doi.org/10.1557/mrs.2016.295.

[17] C. Le Quéré, R.J. Andres, T. Boden, T. Conway, R.A. Houghton, J.I. House, G. Marland, G.P. Peters, G. van der Werf, A. Ahlström, R.M. Andrew, L. Bopp, J.G. Canadell, P. Ciais, S.C. Doney, C. Enright, P. Friedlingstein, C. Huntingford, A.K. Jain, C. Jourdain, E. Kato, R.F. Keeling, K. Klein Goldewijk, S. Levis, P. Levy, M. Lomas, B. Poulter, M.R. Raupach, J. Schwinger, S. Sitch, B.D. Stocker, N. Viovy, S. Zaehle, N. Zeng, The global carbon budget 1959–2011, Earth System Science Data Discussions. 5 (2012) 1107–1157. https://doi.org/info:doi:10.5194/essdd-5-1107-2012.

[18] H. Liu, T. Du, N.M.A. Krishnan, H. Li, M. Bauchy, Topological optimization of cementitious binders: Advances and challenges, Cement and Concrete Composites. (2018). https://doi.org/10.1016/j.cemconcomp.2018.08.002.

[19] G. Constantinides, F.-J. Ulm, The nanogranular nature of C–S–H, Journal of the Mechanics and Physics of Solids. 55 (2007) 64–90. https://doi.org/10.1016/j.jmps.2006.06.003.

[20] M. Vandamme, F.-J. Ulm, P. Fonollosa, Nanogranular packing of C–S–H at substochiometric conditions, Cement and Concrete Research. 40 (2010) 14–26. https://doi.org/10.1016/j.cemconres.2009.09.017.

[21] J. Colombo, E. Del Gado, Stress localization, stiffening, and yielding in a model colloidal gel, Journal of Rheology. 58 (2014) 1089–1116. https://doi.org/10.1122/1.4882021.

[22] K. Ioannidou, R. J.-M. Pellenq, E.D. Gado, Controlling local packing and growth in calcium–silicate–hydrate gels, Soft Matter. 10 (2014) 1121–1133. https://doi.org/10.1039/C3SM52232F.

[23] Ioannidou Katerina, Del Gado Emanuela, Ulm Franz-Josef, Pellenq Roland J.-M., Inhomogeneity in Cement Hydrates: Linking Local Packing to Local Pressure, Journal of Nanomechanics and Micromechanics. 7 (2017) 04017003. https://doi.org/10.1061/(ASCE)NM.2153-5477.0000120.

[24] H.M. Jennings, A model for the microstructure of calcium silicate hydrate in cement paste, Cement and Concrete Research. 30 (2000) 101–116. https://doi.org/10.1016/S0008-8846(99)00209-4.

[25] H. Manzano, E. Masoero, I. Lopez-Arbeloa, H. M. Jennings, Shear deformations in calcium silicate hydrates, Soft Matter. 9 (2013) 7333–7341. https://doi.org/10.1039/C3SM50442E.

[26] H. Manzano, S. Moeini, F. Marinelli, A.C.T. van Duin, F.-J. Ulm, R.J.-M. Pellenq, Confined Water Dissociation in Microporous Defective Silicates: Mechanism, Dipole Distribution, and Impact on Substrate Properties, J. Am. Chem. Soc. 134 (2012) 2208–2215. https://doi.org/10.1021/ja209152n.

[27] R.J.-M. Pellenq, H. Van Damme, Why Does Concrete Set?: The Nature of Cohesion Forces in Hardened Cement-Based Materials, MRS Bulletin. 29 (2004) 319–323. https://doi.org/10.1557/mrs2004.97.

[28] B. Jönsson, A. Nonat, C. Labbez, B. Cabane, H. Wennerström, Controlling the Cohesion of Cement Paste, Langmuir. 21 (2005) 9211–9221. https://doi.org/10.1021/la051048z.

[29] M. Hermes, M. Dijkstra, Jamming of polydisperse hard spheres: The effect of kinetic arrest, EPL. 89 (2010) 38005. https://doi.org/10.1209/0295-5075/89/38005.

[30] G.D. Scott, D.M. Kilgour, The density of random close packing of spheres, J. Phys. D: Appl. Phys. 2 (1969) 863. https://doi.org/10.1088/0022-3727/2/6/311.

[31] A. Donev, I. Cisse, D. Sachs, E.A. Variano, F.H. Stillinger, R. Connelly, S. Torquato, P.M. Chaikin, Improving the Density of Jammed Disordered Packings Using Ellipsoids, Science. 303 (2004) 990–993. https://doi.org/10.1126/science.1093010.

[32] M. Bauchy, Structural, vibrational, and elastic properties of a calcium aluminosilicate glass from molecular dynamics simulations: The role of the potential, The Journal of Chemical Physics. 141 (2014) 024507. https://doi.org/10.1063/1.4886421.

[33] B. Wang, Y. Yu, Y.J. Lee, M. Bauchy, Intrinsic Nano-Ductility of Glasses: The Critical Role of Composition, Front. Mater. 2 (2015) 11. https://doi.org/10.3389/fmats.2015.00011.

[34] J.F. Nye, Physical Properties of Crystals: Their Representation by Tensors and Matrices, Clarendon Press, 1985.

[35] Abdolhosseini Qomi M. J., Ebrahimi D., Bauchy M., Pellenq R., Ulm F.-J., Methodology for Estimation of Nanoscale Hardness via Atomistic Simulations, Journal of Nanomechanics and Micromechanics. 7 (2017) 04017011. https://doi.org/10.1061/(ASCE)NM.2153-5477.0000127.

[36] M. Bauchy, M.J.A. Qomi, C. Bichara, F.-J. Ulm, R.J.-M. Pellenq, Rigidity Transition in Materials: Hardness is Driven by Weak Atomic Constraints, Phys. Rev. Lett. 114 (2015) 125502. https://doi.org/10.1103/PhysRevLett.114.125502.

[37] M.J. Abdolhosseini Qomi, K.J. Krakowiak, M. Bauchy, K.L. Stewart, R. Shahsavari, D. Jagannathan, D.B. Brommer, A. Baronnet, M.J. Buehler, S. Yip, F.-J. Ulm, K.J. Van

Vliet, R.J.-M. Pellenq, Combinatorial molecular optimization of cement hydrates, Nat Commun. 5 (2014) 4960. https://doi.org/10.1038/ncomms5960.

[38] A.P. Thompson, S.J. Plimpton, W. Mattson, General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions, The Journal of Chemical Physics. 131 (2009) 154107. https://doi.org/10.1063/1.3245303.

[39] Y. Yu, M. Wang, N.M. Anoop Krishnan, M.M. Smedskjaer, K. Deenamma Vargheese, J.C. Mauro, M. Balonis, M. Bauchy, Hardness of silicate glasses: Atomic-scale origin of the mixed modifier effect, Journal of Non-Crystalline Solids. 489 (2018) 16–21. https://doi.org/10.1016/j.jnoncrysol.2018.03.015.

[40] B. Wang, N.M.A. Krishnan, Y. Yu, M. Wang, Y. Le Pape, G. Sant, M. Bauchy, Irradiation-induced topological transition in SiO2: Structural signature of networks' rigidity, Journal of Non-Crystalline Solids. 463 (2017) 25–30. https://doi.org/10.1016/j.jnoncrysol.2017.02.017.

[41] Y. Yu, M. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Bauchy, Thermometer Effect: Origin of the Mixed Alkali Effect in Glass Relaxation, Phys. Rev. Lett. 119 (2017) 095501. https://doi.org/10.1103/PhysRevLett.119.095501.

[42] Y. Yu, J.C. Mauro, M. Bauchy, Stretched exponential relaxation of glasses: Origin of the mixed-alkali effect, American Ceramic Society Bulletin. 96 (2017) 34–36.

[43] Y. Yu, M. Wang, D. Zhang, B. Wang, G. Sant, M. Bauchy, Stretched Exponential Relaxation of Glasses at Low Temperature, Phys. Rev. Lett. 115 (2015) 165901. https://doi.org/10.1103/PhysRevLett.115.165901.

[44] M. Vandamme, F.-J. Ulm, Nanogranular origin of concrete creep, PNAS. 106 (2009) 10552–10557. https://doi.org/10.1073/pnas.0901033106.

[45] G.N. Greaves, A.L. Greer, R.S. Lakes, T. Rouxel, Poisson's ratio and modern materials, Nature Materials. 10 (2011) 823–837. https://doi.org/10.1038/nmat3134.

[46] X. Zheng, H. Lee, T.H. Weisgraber, M. Shusteff, J. DeOtte, E.B. Duoss, J.D. Kuntz, M.M. Biener, Q. Ge, J.A. Jackson, S.O. Kucheyev, N.X. Fang, C.M. Spadaccini, Ultralight, ultrastiff mechanical metamaterials, Science. 344 (2014) 1373–1377. https://doi.org/10.1126/science.1252291.

[47] Z. Qin, G.S. Jung, M.J. Kang, M.J. Buehler, The mechanics and design of a lightweight three-dimensional graphene assembly, Science Advances. 3 (2017) e1601536. https://doi.org/10.1126/sciadv.1601536.

[48] L.J. Gibson, M.F. Ashby, The Mechanics of Three-Dimensional Cellular Materials, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 382 (1982) 43–59. https://doi.org/10.1098/rspa.1982.0088.

[49] A.P. Roberts, E.J. Garboczi, Elastic properties of model random three-dimensional open-cell solids, Journal of the Mechanics and Physics of Solids. 50 (2002) 33–55. https://doi.org/10.1016/S0022-5096(01)00056-4.

[50] M.F. Ashby, R.F.M. Medalist, The mechanical properties of cellular solids, Metallurgical Transactions A. 14 (1983) 1755–1769. https://doi.org/10.1007/BF02645546.

[51] K. Philipps, R.P. Stoffel, R. Dronskowski, R. Conradt, Experimental and Theoretical Investigation of the Elastic Moduli of Silicate Glasses and Crystals, Front. Mater. 4 (2017). https://doi.org/10.3389/fmats.2017.00002.

[52] A.L. Greer, Y.Q. Cheng, E. Ma, Shear bands in metallic glasses, Materials Science and Engineering: R: Reports. 74 (2013) 71–132. https://doi.org/10.1016/j.mser.2013.04.001.

[53] P. Schall, D.A. Weitz, F. Spaepen, Structural Rearrangements That Govern Flow in Colloidal Glasses, Science. 318 (2007) 1895–1899. https://doi.org/10.1126/science.1149308.

[54] I. Pignatelli, A. Kumar, R. Alizadeh, Y. Le Pape, M. Bauchy, G. Sant, A dissolution-precipitation mechanism is at the origin of concrete creep in moist environments, The Journal of Chemical Physics. 145 (2016) 054701. https://doi.org/10.1063/1.4955429.

[55] M. Bauchy, M. Wang, Y. Yu, B. Wang, N.M.A. Krishnan, E. Masoero, F.-J. Ulm, R. Pellenq, Topological Control on the Structural Relaxation of Atomic Networks under Stress, Physical Review Letters. 119 (2017). https://doi.org/10.1103/PhysRevLett.119.035502.

Section A. Physics-Driven Computational Simulations: Make the Physics Simple

A3. Accelerated Molecular Dynamics Simulation

Chapter 4. Long-Term Creep Deformations in Colloidal Calcium–Silicate–Hydrate Gels by Accelerated Aging Simulations

4.1 Introduction

Jammed colloidal gels—i.e., aggregated systems made of interacting nanograins [1,2]—are widely used in many industrial fields [3–5]. When subjected to a sustained load, jammed colloidal gels can feature some delayed viscoplastic creep deformations that can ultimately result in macroscopic failure [6–8]. Specifically, creep deformations in calcium–silicate–hydrate (C–S–H) gels—the glue of concrete that forms upon the hydration of cement [5,9,10]—can decrease the lifespan of concrete structures [11–15]. This is significant as the maintenance or replacement of structures impacted by creep involves the use of large quantities of cement and concrete, which come with a significant environmental burden [6,16–18]. As such, the prediction of long-term creep deformations in C–S–H (and colloidal gels in general) could facilitate the design of new binders featuring minimal creep.

However, although various models have been proposed to explain the origin of concrete creep [11,15,19–22], the prediction of long-term creep deformations remains challenging. This arises from the facts that (i) cement binders are complex, multi-scale materials [5,9,23], (ii) various scales (atomic, mesoscale, etc.) may contribute to controlling creep [12], and (iii) creep deformations are associated with extended timescales, which far exceed the timescale accessible to conventional computional simulation methods (e.g., molecular dynamics or coarse-grained mesoscale simulations) [6,24,25].

To overcome the timescale limitation of conventional physics-based simulations techniques, we recently showed that stress perturbations cycles can be efficiently used to accelerate the aging of disordered, out-of-equilibrium materials [6,24,26]. Here, building on these ideas, we

report some accelerated simulations of creep deformations in C–S–H based on the mesoscale model introduced by Masoero *et al.* [5]. We obtain a very good agreement with nanoindentation creep tests, which suggests that the reorganization of C–S–H grains at the mesoscale controls the creep of concrete. Based on these results, we show that the creep of C–S–H increases logarithmically with time, which is in line with experimental results from nanoindentation and with the predictions from the free-volume dynamics theory of granular physics [11,27]. Further, we establish the existence of a linear regime wherein creep deformations linearly depend on the applied load, which allows us to define a "creep modulus" material constant. These findings could offer a new physics-based basis for nanoengineering colloidal gels featuring minimal creep.

This paper is organized as follows. In Sec. 4.2, we describe the methodology used herein to generate the C–S–H mesoscale configurations and model their creep deformations. In Sec. 4.3, we validate out simulations based on nanoindentation data and investigate the nature of creep deformations in C–S–H. Some consequences in the mechanism of creep in C–S–H are discussed in Sec. 4.4. Finally, some conclusions are presented in Sec. 4.5.

4.2 Methods

4.2.1 Preparation of the C–S–H configurations

We adopt here the colloidal model of C–S–H introduced by Masoero *et al.* [5,28], as it has been found to offer a realistic description of the mesoscale structure and nanomechanics of C–S–H [5,28,29]. In this model, the C–S–H gel is described as an ensemble of polydisperse spherical grains that interact with each other via a generalized Lennard-Jones interaction energy potential:

$$U_{ij}(r_{ij}) = 4\varepsilon(\sigma_i, \sigma_j)\left[\left(\frac{\bar{\sigma}_{ij}}{r_{ij}}\right)^{2\alpha} - \left(\frac{\bar{\sigma}_{ij}}{r_{ij}}\right)^{\alpha}\right] \qquad \textbf{Eq. (4-1)}$$

where σ_i and σ_j are the diameters of grains i and j, $\bar{\sigma}_{ij} = (\sigma_i + \sigma_j)/2$ is the average diameter for a given pair of atom, α is a parameter that controls the narrowness of the potential well, r_{ij} is distance between the centers of the grains i and j, and $\varepsilon(\sigma_i, \sigma_j)$ is the depth of the potential energy well. By considering each pair of grains in contact as two springs in series, the depth is given by $\varepsilon(\sigma_i, \sigma_j) = A_0 \beta_{ij} {\bar{\sigma}_{ij}}^3$, where $A_0 = kE$ is a prefactor that is proportional to the bulk Young's modulus E of a grain, wherein $k = 0.002324$ (computed by the serial spring model) and $E = 63.6$ GPa (based on previous atomistic simulations of bulk C–S–H) [28,30]. $\beta_{ij} = \sigma_i \sigma_j / {\bar{\sigma}_{ij}}^2$ is a correction term arising from the serial arrangement. The potential defined in Eq. (4-1) shows a minimum at $r_m = \sqrt[\alpha]{2}\bar{\sigma}_{ij}$ so that the effective diameter of a grain i is defined as $\sigma_{0,i} = \sqrt[\alpha]{2}\sigma_i$. The attractive force is maximum at a distance $r_u = \sqrt[\alpha]{\frac{4\alpha+2}{\alpha+1}}\bar{\sigma}_{ij}$ so that, by choosing $\alpha = 14$, the tensile strain at failure $\varepsilon_u = (r_u - r_m)/r_m$ is close to the value of 5% obtained in previous atomistic simulation of bulk C–S–H [28,30,31].

The C–S–H configurations are generated by grand canonical Monte Carlo (GCMC) simulations, as described in the following [5,29,32]. Starting from an initially empty cubic box with a size ranging from 600 to 920 Å, some C–S–H grains are iteratively inserted, wherein the size of each grain is randomly selected from a uniform distribution between a minimum σ_m and a maximum σ_M value. Experimentally, the polydispersity of the C–S–H grains is strongly supported by the absence of a clear characteristic size in SANS neutron scattering [9]. The standard deviation θ of the distribution is then used to define the polydispersity index of the configuration as: [5,28]

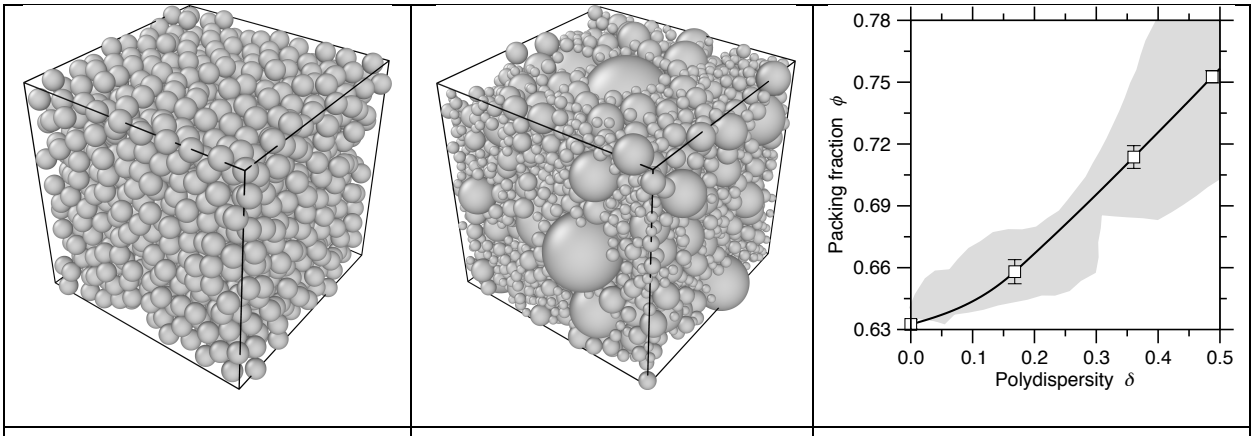
$$\delta = \theta/[(\sigma_m + \sigma_M)/2] \qquad \textbf{Eq. (4-2)}$$

Here, various polydispersity values are considered, with σ ranging from 3.0 to 35 nm, and the number of grains at saturation ranging from 1700 to 7000. In detail, each GCMC step comprises

5 attempts of grain insertions or deletions followed by 500 attempts to randomly displace an existing grain. At each step, the probability of acceptance of the attempt is given by $\min\{1,\ \exp[-(\Delta U - \mu\Delta N)/k_{\mathrm{B}}T]\}$ [33], where k_{B} is the Boltzmann constant, T the temperature, ΔU the variation in potential energy caused by the trial move, ΔN the variation in the number of C–S–H grains, and μ the chemical potential, which is taken here as $-2k_{\mathrm{B}}T$ based on previous studies [34]. This value ensures the formation of a realistic final structure within a reasonable simulation time. Note that, here, the chemical potential does not bear a quantitative meaning and that small variation in the chemical potential do not significantly alter the structure and properties of the simulated C–S–H samples [34,35]. This process is iteratively repeated until the number of inserted grains reaches a plateau. Note that the GCMC process is performed at constant volume—so that some tensile pressure builds up in the system upon precipitation. At the end of the GCMC simulation, such pressure is released by subjecting the system to a molecular dynamics relaxation in the *NPT* ensemble at zero stress, eventually followed by a final energy minimization. The packing fraction ϕ of each configuration is then computed as $\phi = [\sum_i((\pi/6){\sigma_{0,i}}^3)]/V$, where V is the volume of the simulation box. Note that five independent simulations of C–S–H precipitation are performed for each degree of polydispersity to calculate the mean value and standard deviation of all the properties presented in the following.

In agreement with previous simulations [5,36], we observe that the C–S–H models that exhibit higher degrees of polydispersity eventually reach higher final packing fraction values—as small grains are able to fill the space left in between larger grains (see Fig. 4-1). For monodisperse configurations, the packing fraction at saturation is found to be around 0.63, that is, close to the theoretical packing limit of random monodisperse spheres [37,38]. Further, we note that the

evolution of the packing fraction at saturation with polydispersity is in good agreement with the range of data observed in previous simulations [5,36].



(a) **(b)** **(c)**

Figure 4-1: Snapshots of select **(a)** monodisperse (polydispersity index $\delta = 0$, see Eq. (4-2)) and **(b)** polydisperse C–S–H gel configurations ($\delta = 0.49$). **(c)** Computed packing fraction (at saturation) of the C–S–H gel configurations as a function of the polydispersity index. The grey area indicates the range of data observed in previous simulations [5,36].

4.2.2 Accelerated aging simulation methodology

We now focus on the methodology introduced herein to simulate creep. As mentioned above, the long-term nature of creep deformations far exceeds the typical timescale accessible to (coarse-grained) molecular dynamics simulations (i.e., from nano- to microseconds at most). Although kinetic Monte Carlo simulations could, in theory, describe the dynamics of the system over up to a few seconds, the application of this technique to polydisperse colloidal gels is challenging due to the high mobility of the small grains—which results in the existence of a large number of small energy barriers [25]. As such, the direct simulation of the stress-induced creep deformation dynamics of C–S–H is, at this point, unachievable.

To overcome this limitation, we present here an accelerated simulation technique that is inspired by previous studies focusing on the relaxation of disordered atomic networks [6,24,26]. Refs. [24,26] provide some technical details on our accelerated method and offer an enthalpy landscape interpretation to the acceleration in the system dynamics that our technique yields. This technique relies on the application of small stress perturbations, which can accelerate the relaxation of out-of-equilibrium materials. Here, to simulate creep under sustained deviatoric load, the mesoscale C–S–H configurations are subjected an average shear stress τ_0 combined with small, cyclic perturbations of shear stress $\pm\Delta\tau$ (see Fig. 4-2). At each stress cycle, a minimization of the energy is performed, with the system having the ability to deform (shape and volume) in order to reach the target stress.

This method is inspired by the artificial aging and rejuvenation experienced by granular materials subjected to vibrations [39]. Namely, small vibrations can induce a compaction of granular materials, making the system *overage*. In contrast, large vibrations tend to randomize the grain configuration, which decreases the overall compactness and, therefore, makes the system *rejuvenate*. Similar ideas, relying on the energy landscape framework [7,8], have been applied to amorphous solids—based on the idea that an external stress tends to deform the energy landscape locally explored by the atoms. The application of a small external stress can result in the removal of some energy barriers existing at zero stress, thereby allowing some atoms to jump toward a new energy basin and relax to lower energy states. This transformation is irreversible as, once the stress is removed, the system remains in its aged state. In contrast, the application of a large stress can make the system move far from its initial state, which eventually leads to rejuvenation—i.e., similar to thermal annealing [40]. As such, a succession of many of such small stress perturbations

can be used to simulate the delayed relaxation of a disordered configuration subjected to a sustained load, i.e., creep.

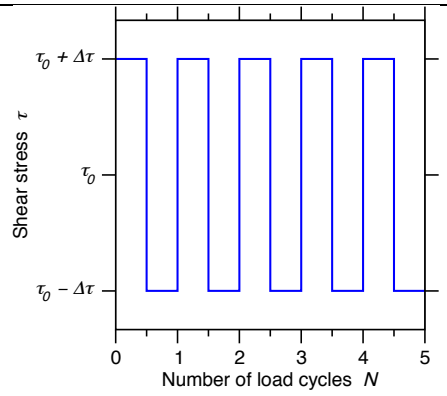


Figure 4-2: Schematic presenting the stress perturbation cycles applied during our accelerated aging simulation method, where τ_0 is the average shear stress (i.e., causing the creep deformation of the colloidal C–S–H gel) and $\Delta\tau$ is the amplitude of the stress perturbations.

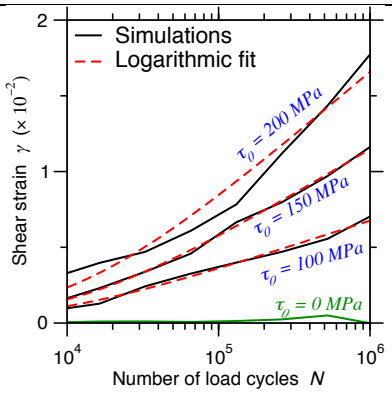


Figure 4-3: Computed shear strain in monodisperse colloidal C–S–H gels with respect to the number of stress perturbation cycles N and for select average shear stress τ_0 values. The amplitude of the stress perturbations $\Delta\tau$ is here set as 30 MPa. The dashed lines are some logarithmic fits following Eq. (4-3).

4.3 Results

4.3.1 Logarithmic nature of creep in C–S–H

Figure 4-3 shows the evolution of the shear strain γ of select C–S–H systems—under a sustained shear stress τ_0—obtained using our accelerated simulation method. Overall, we observe that the application of the stress cycles results in a gradual increase in shear strain. Note that, at zero average shear stress (i.e. $\tau_0 = 0$ MPa) no noticeable shear strain occurs. Further, we note that γ increases logarithmically with the number of applied stress cycles N and linearly with the applied stress τ_0. This suggests that the shear strain induced by creep can be expressed as:

$$\gamma(N) = (\tau_0/C)\log(1 + N/N_0) \qquad \textbf{Eq. (4-3)}$$

where C is the creep modulus and N_0 a fitting parameter that is analogous to a relaxation time [6,11]. Note that the number of load cycles N has been demonstrated to be equivalent to a fictitious time t, that is, $t = N\Delta t$, wherein Δt is a constant duration corresponding to each stress cycle [6,24,26]. This arises from the fact that the height of the energy barriers through which the system transits across each cycle, remains roughly constant over successive cycles [6]. As such, on the basis of transition state theory, the time needed for a system to jump over an energy barrier E_A is constant and proportional to $\exp(-E_A/kT)$, where k is the Boltzmann constant and T the temperature [41]. However, the fictitious time associated with each stress cycle cannot be directly mapped into real time. In other words, we cannot ensure the trajectory of the C–S–H grains predicted by our accelerated simulation method is fully equivalent to the one that would be observed upon spontaneous creep. However, we previously demonstrated that macroscopic properties (e.g., strain), which are not very sensitive to the microscopic details of the system, exhibit a realistic evolution with the fictitious time [6]. The logarithmic nature of C–S–H creep observed herein is in good agreement with nanoindentation data [11,42] and such a logarithmic evolution has also

been observed in the creep of various materials [43,44]. Similarly, a logarithmic compaction is also found in granular materials that are subjected to vibrations [39].

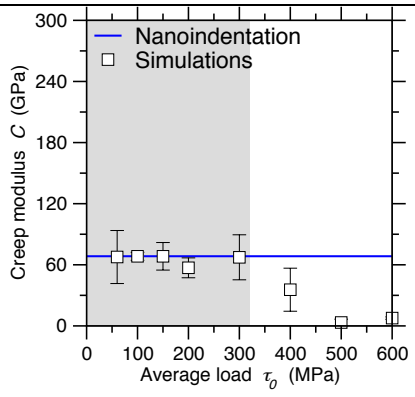
4.3.2 Linear regime and creep modulus as a material constant

We now focus on the dependence of the shear strain γ on the applied shear stress τ_0. As expected, we observe that the magnitude of the creep-induced deformation increases with increasing values of applied load (see Fig. 4-3). The relationship between γ and τ_0 is effectively captured by the value of the creep modulus *C*—which should be constant if γ increases linearly with τ_0. Note that monodisperse C–S–H configurations and small stress perturbation values (here taken as 30 MPa) are first considered in this section.

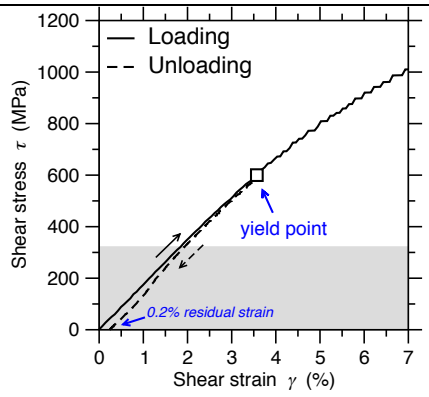
Figure 4-4a presents the evolution of the creep modulus, which is obtained by fitting the strain curves such as those presented in Fig. 4-3 by the logarithmic law given in Eq. (4-3). Interestingly, we observe that, at low τ_0 values, the value of the creep modulus is constant and does not depend on τ_0. However, we note that the value of the creep modulus drastically decreases once the applied shear stress τ_0 exceeds a critical value (which is here found to be around 320 MPa). Note that, in the low-stress regime, the N_0 constant is also found to be constant, which indicates that the mapping between number of stress cycles and corresponding creep time does not depend on the applied load.

Importantly, the fact that the creep modulus exhibits a constant value upon the application of low loads suggests that, under this regime, creep deformations feature a linear dependence on the applied load (see Eq. (4-3)), in agreement with nanoindentation data [11,42]. This observation also establishes the creep modulus as an intrinsic material constant, that is, that only depends on the material composition and structure [6]. In addition, the linear nature of creep observed herein

strongly supports the fact that, despite the large difference in length and time scales, small-scale creep deformations obtained by nanoindentation (obtained over a few seconds) should yield similar values of creep modulus than macroscopic creep tests (obtained over much longer periods of time).



(a)



(b)

Figure 4-4: (a) Computed creep modulus C (see Eq. (4-3)) in monodisperse colloidal C–S–H gels under varying average shear stress values τ_0 (for $\Delta\tau$ = 30 MPa). The results are compared with experimental nanoindentation data (blue line) [11,42]. **(b)** Computed stress–strain curve in a monodisperse colloidal C–S–H gel upon shearing. The dash line is an unloading curve from the yield point (square point), where the residual strain is found to be 0.2%. In both panels, the grey window shows the range of average load values τ_0 wherein creep is linear.

4.3.3 Limits of the linear regime

We now investigate the origin of the departure from the linear regime at large stress (see Fig. 4-4a). To this end, Fig. 4-4b shows stress–strain behavior of monodisperse C–S–H under shear—wherein the C–S–H configuration is subjected to a pure shear deformation by gradually increasing the shear stress and performing an energy minimization after each increment of stress. As expected, we observe that, at low stress, shear stress increases linearly with shear strain, which

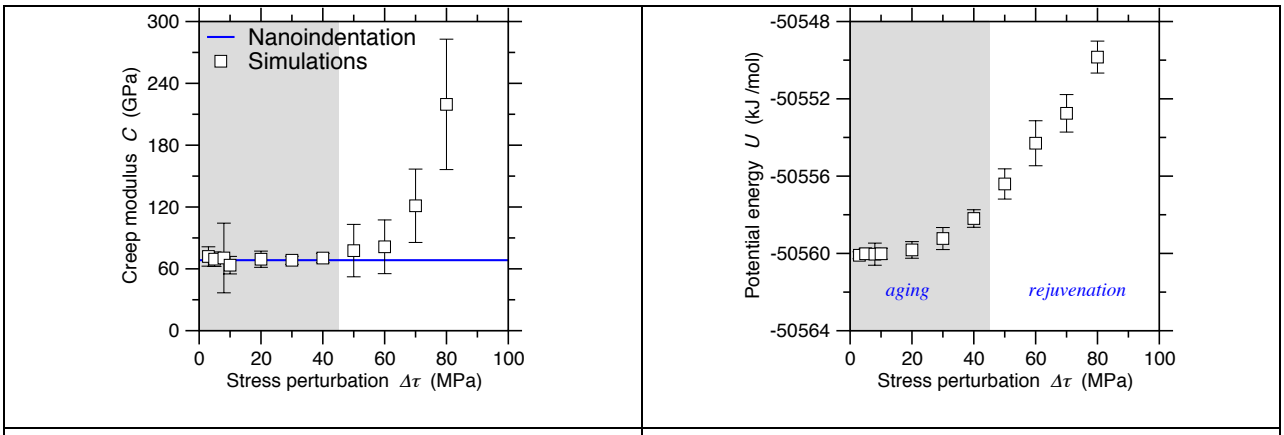
characterizes a linear elastic deformation. We note that the system starts to exhibit some yielding around 600 MPa, which manifests itself by a deviation from linearity in the stress–strain curve and the existence of a residual permanent strain upon unloading. Based on this result, we conclude that creep remains linear as long as the applied load remains low as compared to the yield point of the material. These results echo the conclusions of a previous study, wherein it was found that a mathematical condition to have a constant creep modulus C is that the activation energy for the irreversible rearrangements increases as the logarithm of the shear strain—a condition that was found to be valid only when the applied stress is lower than the yield stress [45]. Here, the present results suggest that the linear regime of creep extends up to stress values that are about 60% of the yield stress threshold. This can be understood from the fact that, when the load approaches the yield point, the material starts to experience local yielding, which results in a drop in creep modulus—i.e., a drastic increase in creep compliance (see Fig. 4-4a).

4.3.4 Aging and rejuvenation in C–S–H under stress perturbations

We now assess the influence of the amplitude of the stress perturbations used herein to accelerate the dynamics of C–S–H under creep. Figure 4-5a shows the creep modulus value C (computed under a constant shear stress of 100 MPa, i.e., in the linear regime) as a function of the amplitude of the stress perturbation $\Delta\tau$. We observe that, for low values of $\Delta\tau$, the obtained creep modulus remains largely constant. This indicates that, in this regime, the creep modulus value yielded by our methodology is not affected by the specific choice of $\Delta\tau$—which is an important observation that confirms the reliability of our approach.

However, we observe that C suddenly increases when $\Delta\tau$ becomes larger than a threshold value (found to be around 45 MPa herein). This indicates that the accelerated creep of the system

is only achieved over a certain range of stress perturbation magnitude (i.e., less than 45 MPa). In contrast, larger values of stress perturbation amplitude do not result in any significant creep deformation, which manifests itself by an increase in C—i.e., an increase in the apparent resistance to creep under fixed external load.

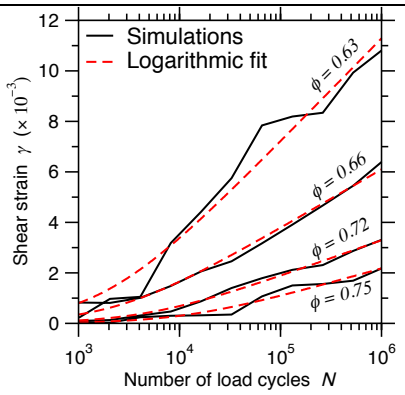


(a) **(b)**

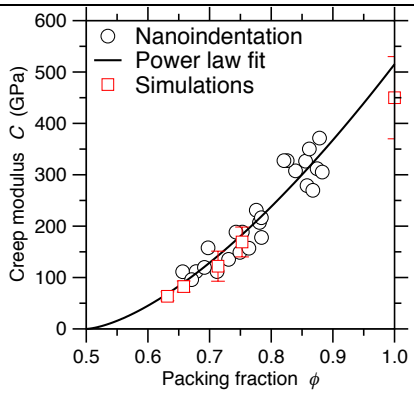
Figure 4-5: (a) Computed creep modulus C (see Eq. (4-3)) in monodisperse colloidal C–S–H gels under varying stress perturbation amplitudes $\Delta\tau$ (for τ_0 = 100 MPa). The results are compared with experimental nanoindentation data (blue line) [11,42]. **(b)** Computed molar potential energy of monodisperse colloidal C–S–H gels at fixed shear strain deformation (γ = 0.2%) under select stress perturbation amplitudes $\Delta\tau$ (for τ_0 = 100 MPa). In both panels, the grey window shows the range of $\Delta\tau$ values wherein C is constant.

This observation can be understood as a balance between stress-induced overaging and rejuvenation. Indeed, as mentioned above, it has been observed that the application of a small stress tends to make a system overage (i.e., accelerate the spontaneous aging of an out-of-equilibrium system) by deforming the energy landscape and suppressing some preexisting energy barriers [40,46]. In turn, the application of a large stress can induce some rejuvenation by significantly moving the system away from its initial position in the energy landscape [8,40,46].

To demonstrate the effect, we compute the potential energy U of a monodisperse C–S–H system having experienced upon creep a constant shear strain deformation ($\gamma = 0.2\%$). Figure 4-5b shows the evolution of U as a function of the stress perturbation amplitude $\Delta\tau$ used to stimulate creep. We observe that, for small values of $\Delta\tau$ (i.e. less than 45 MPa), the energy of the system remains fairly independent of $\Delta\tau$. In contrast, for larger values of $\Delta\tau$, we observe a significant increase in U. This confirms that large values of stress perturbations amplitude results in a rejuvenation of the system, that is, a destabilization toward higher energy states. This *a posteriori* confirms that the energy state of the system experiencing creep is not affected by the choice of $\Delta\tau$ as long as no rejuvenation is induced.



(a)



(b)

Figure 4-6: (a) Computed shear strain in polydisperse colloidal C–S–H gels with respect to the number of stress perturbation cycles and for select packing densities ϕ (with $\tau_0 = 100$ MPa and $\Delta\tau = 30$ MPa). The dashed lines are some logarithmic fits (see Eq. (4-3)). **(b)** Computed creep modulus C (see Eq. (4-3)) of C–S–H as a function of the packing fraction. The results are compared with experimental nanoindentation data [11] and data from a previous atomistic simulation of creep in bulk C–S–H at $\phi = 1$, that is, with no porosity [6]. The solid line is a power-law fit (see Eq. (4-4)).

4.3.5 Experimental validation of our accelerated simulation technique

Having established the range of $\Delta\tau$ values for which our methodology yield an accelerated creep dynamics without inducing any rejuvenation, we are now in position to compare our computed results with those obtained experimentally. Note that such a direct comparison may be challenging due to the fact that the state of stress experienced in experiments (e.g., nanoindentation) can significantly differ from that imposed herein. Nevertheless, the fact that (i) the value of the creep modulus and (ii) the nature of the mapping between number of stress cycles and corresponding creep time both do not depend on the applied load (see Sec. 4.3.2) makes it possible to meaningfully compare computed and experimental data.

Based on these observations, we now assess the effect of the packing density of C–S–H on its creep modulus and compare the outcome to nanoindentation data [11,42]. The nanoindentation experiments—whose outcomes are used herein to validate our simulations—were conducted on cementitious binders formed upon the hydration of ordinary portland cement [11]. The creep modulus was determined by applying a constant load and measuring the time-dependent displacement of the indenter. A clustering algorithm was used to isolate the properties of C–S–H from those of the other phases [11]. Figure 4-6a shows the shear strain exhibited by C–S–H configurations for select packing density values. Overall, we observe the conclusions previously established in the case of monodisperse C–S–H are retained for polydisperse systems, namely, (i) creep exhibits a logarithmic dependence on time, (ii) creep is load-linear as long as the applied stress remains smaller than the yield stress of the system, and (iii) the value of the computed creep modulus is independent of the amplitude of the stress perturbations as long as no rejuvenation is induced. This allows us to compute the evolution of the creep modulus as a function of packing density following Eq. (4-3). As shown in Fig. 4-6b, we observe that the creep modulus increases

with increasing values of C–S–H packing density. Importantly, we obtain an excellent agreement between simulation and nanoindentation data [11,42], which strongly supports the ability of our model and accelerated simulation method to offer a realistic description of the creep of C–S–H.

4.4 Discussion

The good agreement between the creep modulus data presented in Fig. 4-6b with nanoindentation suggests that the mesoscale model of C–S–H is able to properly capture the mechanism of creep in C–S–H and, in turn, that other features that are not considered by the present model do not necessarily need to be accounted for to model C–S–H creep (see below). This shows that, under load, the creep of C–S–H occurs via some structural reorganization within the mesoscale structure of C–S–H. Specifically, the fact that our mesoscale model of C–S–H is based on grains of constant geometry suggests that creep does not arise from time-dependent variations in the volume or shape of the C–S–H grains, but rather arise from some reorganization in the mesoscale structure of the grains. Nevertheless, such mesoscale rearrangements necessarily occur through some atomic-scale deformations at the interface between C–S–H grains. Then, the fact that the present mesoscale model yields some creep modulus values that are in good agreement with experiments although it only considers spherical grains suggests the shape of the grains may not have a first-order effect on the height of the energy barriers that need to be overcome upon creep. In addition, since the present mesoscale model relies on a spherical, isotropic description of the C–S–H grains, our results suggest that the relative orientation of the C–S–H grains with respect to each other may not affect creep to the first order—which may arise from the fact that the local packing density has a first order effect in controlling the magnitude of the energy barriers in such dense systems. Finally, the fact that our model does not incorporate any transversal force opposing

the sliding among particles suggests that frictional and shearing effects acting the level of the interlayer space in the C–S–H grains are not necessarily the sole or main factor responsible for creep of C–S–H, which in agreement with previous studies [47].

Overall, our results suggest that the nature of C–S–H creep is primarily nanogranular. This is also supported by the logarithmic time-evolution of creep (see Sec. 4.3.1), which can elegantly be explained by the free-volume theory (FVT) in granular physics. In details, FVT assumes that the creep/aging of the system occurs via some structural reorganization of the grains that jump into some local free space—so that the changing rate of the packing density $\dot{\phi}$ (and, hence, the creep deformation rate $\dot{\gamma}$) is expected to be proportional to the amount of holes that are larger than the grain size. It is then assumed that that size of the holes (i.e., local free volume Ω) exhibits a Poisson's distribution $p(\Omega) = \frac{1}{\theta}\exp(-\Omega/\theta)$, wherein θ the average free volume per grain. The normalized amount of accessible holes per jumping grain can be expressed as the probability of holes that are larger than the grain size (ρ), that is, $\int_{\rho}^{\infty} p(\Omega)\, d\Omega = \exp(-\rho/\theta)$. As such, FVT predicts that the deformation rate $\dot{\phi}$ decreases exponentially as a function of the excluded volume θ, i.e., $\dot{\phi} \propto \exp(-\rho/\theta)$. Further, one gets $\theta = \rho(1/\phi - 1/\phi_{\infty}) \approx \rho{\phi_{\infty}}^{-2}(\phi_{\infty} - \phi)$, wherein ϕ represents the packing density of current aging system and ϕ_{∞} is the ultimately limit packing density (note that, at the vicinity of the jamming threshold, $|\phi_{\infty} - \phi| \ll 1$). Finally, the time integration of $\dot{\phi} \propto \exp(-\rho/\theta)$ yields $t \propto \exp(\rho/\theta)$, so that $\dot{\phi} \propto t^{-1}$ [11,27].

The nanogranular nature of C–S–H's creep is also supported by the dependence of the creep modulus on the packing density (see Fig. 4-6b). In details, we find that the evolution of the creep modulus can be well described by a power law: [11,42]

$$C = C_0 \times (2\phi - 1)^{\alpha} \qquad \textbf{Eq. (4-4)}$$

where α is the power law exponent and C_0 the value of the creep modulus at zero porosity (i.e., $\phi = 1$). This trend can be understood from the fact that C–S–H must exhibit a minimum packing fraction that is larger than 0.5 to present a percolated structure featuring a non-zero stiffness, hardness, and resistance to creep [9,29]. Starting from this threshold, the increase in C (i.e., increase in C–S–H's resistance to creep) with increasing packing density is consistent with the free-volume theory framework [27]—since the level of free space (θ) accessible to the C–S–H grains decreases upon increasing packing density ϕ. In details, the relationship:

$$\dot{\phi} \propto \exp(-\rho/\theta) = \exp\left[-(1/\phi - 1/\phi_\infty)^{-1}\right] \qquad \textbf{Eq. (4-5)}$$

established above based on the FVT framework suggests that the deformation rate $\dot{\phi}$ decreases (and, therefore, that the resistance to creep C increases) with increasing packing density. From a mechanism viewpoint, this trend suggests that, by filling the existing free space left in the mesostructure, the presence of small C–S–H grains in between the larger C–S–H grains effectively reduces the number of possible structural reorganizations and, thereby, reduces the propensity for creep. Finally, it should be pointed out that, although our results are consistent with free-volume theory, our simulations do not exclude other possible physical origins. Clearly, more work is needed to further investigate the nature of the structural reorganization occurring during creep.

Note that the present results do not involve that the atomic structure and composition of the C–S–H grains is irrelevant. Actually, these atomic details are already embedded in the effective potential energy governing the mutual interactions between each pair of C–S–H grains (following Eq. (4-1)). Namely, a variation in the atomic composition and structure of the C–S–H grains would affect the values of the parameters used in Eq. (4-1), which, in turn, would change the dynamics of creep. In fact, the role of the atomic scale is captured in the C_0 term in Eq. (4-4), that is, the creep modulus at zero porosity (i.e., in the absence of any free volume). Notably, the creep

modulus at zero porosity that can here be obtained by extrapolating our computed results toward $\phi = 1$ is in very good agreement with the data predicted by atomic-scale molecular dynamics simulations (see Fig. 4-6b) [6,15]. This highlights the importance of both the atomic and mesoscale structure in controlling the creep of cementitious binders.

4.5 Conclusions

To summarize, by accelerating the aging of the C–S–H mesostructure when subjected to a sustained load, our accelerated simulation method can properly describe the long-term creep of C–S–H and yield a quantitative agreement with experimental nanoindentation data. We observe that the creep of C–S–H exhibit a logarithmic dependence on time and a linear dependence on the applied load—which supports the nanogranular nature of C–S–H creep. Importantly, this work offers the first consistent description of the effect of packing on the propensity for creep of C–S–H. Our modeling framework now makes it possible to further investigate the structural mechanism of creep in C–S–H gels, explore the potential for the discovery of creep-resistant structures, and investigate the effect of each of the hypothesis/parameters of our model. From a practical viewpoint, our methodology can be used to predict the long-term creep deformation of C–S–H gels, which is challenging to access experimentally due to associated time and length scales [11,42]. More generally, our accelerated aging methodology is generic and can be applied to investigate the mechanism(s) governing the relaxation of out-of-equilibrium phases, e.g., glassy, colloidal, or granular materials.

4.6 References

[1] V. Trappe, V. Prasad, L. Cipelletti, P.N. Segre, D.A. Weitz, Jamming phase diagram for attractive particles, Nature. 411 (2001) 772–775. https://doi.org/10.1038/35081021.

[2] M.Y. Lin, H.M. Lindsay, D.A. Weitz, R.C. Ball, R. Klein, P. Meakin, Universality in colloid aggregation, Nature. 339 (1989) 360–362. https://doi.org/10.1038/339360a0.

[3] Y.M. Joshi, Dynamics of Colloidal Glasses and Gels, Annual Review of Chemical and Biomolecular Engineering. 5 (2014) 181–202. https://doi.org/10.1146/annurev-chembioeng-060713-040230.

[4] Wang Q., Wang L., Detamore M. S., Berkland C., Biodegradable Colloidal Gels as Moldable Tissue Engineering Scaffolds, Advanced Materials. 20 (2007) 236–239. https://doi.org/10.1002/adma.200702099.

[5] E. Masoero, E. Del Gado, R.J.-M. Pellenq, F.-J. Ulm, S. Yip, Nanostructure and Nanomechanics of Cement: Polydisperse Colloidal Packing, Phys. Rev. Lett. 109 (2012) 155503. https://doi.org/10.1103/PhysRevLett.109.155503.

[6] M. Bauchy, M. Wang, Y. Yu, B. Wang, N.M.A. Krishnan, E. Masoero, F.-J. Ulm, R. Pellenq, Topological Control on the Structural Relaxation of Atomic Networks under Stress, Physical Review Letters. 119 (2017). https://doi.org/10.1103/PhysRevLett.119.035502.

[7] D.J. Lacks, Energy Landscapes and the Non-Newtonian Viscosity of Liquids and Glasses, Physical Review Letters. 87 (2001). https://doi.org/10.1103/PhysRevLett.87.225502.

[8] D.J. Lacks, M.J. Osborne, Energy Landscape Picture of Overaging and Rejuvenation in a Sheared Glass, Physical Review Letters. 93 (2004). https://doi.org/10.1103/PhysRevLett.93.255501.

[9] K. Ioannidou, K.J. Krakowiak, M. Bauchy, C.G. Hoover, E. Masoero, S. Yip, F.-J. Ulm, P. Levitz, R.J.-M. Pellenq, E.D. Gado, Mesoscale texture of cement hydrates, PNAS. 113 (2016) 2029–2034. https://doi.org/10.1073/pnas.1520487113.

[10] H.M. Jennings, Refinements to colloid model of C-S-H in cement: CM-II, Cement and Concrete Research. 38 (2008) 275–289. https://doi.org/10.1016/j.cemconres.2007.10.006.

[11] M. Vandamme, F.-J. Ulm, Nanogranular origin of concrete creep, PNAS. 106 (2009) 10552–10557. https://doi.org/10.1073/pnas.0901033106.

[12] Z. Bažant, M. Hubler, R. Wendner, Q. Yu, Progress in Creep and Shrinkage Prediction Engendered by Alarming Bridge Observations and Expansion of Laboratory Database, in: Mechanics and Physics of Creep, Shrinkage, and Durability of Concrete, American Society of Civil Engineers, 2013: pp. 1–17.

http://ascelibrary.org/doi/abs/10.1061/9780784413111.001 (accessed December 18, 2014).

[13] Z.P. Bažant, M.H. Hubler, Q. Yu, Excessive creep deflections: An awakening, Concrete International. 33 (2011) 44–46.

[14] H.M. Jennings, Colloid model of C−S−H and implications to the problem of creep and shrinkage, Mat. Struct. 37 (2004) 59–70. https://doi.org/10.1007/BF02481627.

[15] I. Pignatelli, A. Kumar, R. Alizadeh, Y.L. Pape, M. Bauchy, G. Sant, A dissolution-precipitation mechanism is at the origin of concrete creep in moist environments, The Journal of Chemical Physics. 145 (2016) 054701. https://doi.org/10.1063/1.4955429.

[16] H. Liu, T. Du, N.M.A. Krishnan, H. Li, M. Bauchy, Topological optimization of cementitious binders: Advances and challenges, Cement and Concrete Composites. (2018). https://doi.org/10.1016/j.cemconcomp.2018.08.002.

[17] C. Le Quéré, R.J. Andres, T. Boden, T. Conway, R.A. Houghton, J.I. House, G. Marland, G.P. Peters, G. van der Werf, A. Ahlström, R.M. Andrew, L. Bopp, J.G. Canadell, P. Ciais, S.C. Doney, C. Enright, P. Friedlingstein, C. Huntingford, A.K. Jain, C. Jourdain, E. Kato, R.F. Keeling, K. Klein Goldewijk, S. Levis, P. Levy, M. Lomas, B. Poulter, M.R. Raupach, J. Schwinger, S. Sitch, B.D. Stocker, N. Viovy, S. Zaehle, N. Zeng, The global carbon budget 1959–2011, Earth System Science Data Discussions. 5 (2012) 1107–1157. https://doi.org/info:doi:10.5194/essdd-5-1107-2012.

[18] M. Bauchy, Nanoengineering of concrete via topological constraint theory, MRS Bulletin. 42 (2017) 50–54. https://doi.org/10.1557/mrs.2016.295.

[19] Z. Bažant, S. Prasannan, Solidification Theory for Concrete Creep. I: Formulation, Journal of Engineering Mechanics. 115 (1989) 1691–1703. https://doi.org/10.1061/(ASCE)0733-9399(1989)115:8(1691).

[20] Z. Bažant, A. Hauggaard, S. Baweja, F. Ulm, Microprestress-Solidification Theory for Concrete Creep. I: Aging and Drying Effects, Journal of Engineering Mechanics. 123 (1997) 1188–1194. https://doi.org/10.1061/(ASCE)0733-9399(1997)123:11(1188).

[21] R. Sinko, M. Vandamme, Z.P. Bažant, S. Keten, Transient effects of drying creep in nanoporous solids: understanding the effects of nanoscale energy barriers, Proc. R. Soc. A. 472 (2016) 20160490. https://doi.org/10.1098/rspa.2016.0490.

[22] A. Morshedifard, S. Masoumi, M.J.A. Qomi, Nanoscale origins of creep in calcium silicate hydrates, Nature Communications. 9 (2018) 1785. https://doi.org/10.1038/s41467-018-04174-z.

[23] M.J.A. Qomi, E. Masoero, M. Bauchy, F.-J. Ulm, E.D. Gado, R.J.-M. Pellenq, C-S-H across Length Scales: From Nano to Micron, in: CONCREEP 10, American Society of Civil Engineers, n.d.: pp. 39–48. http://ascelibrary.org/doi/abs/10.1061/9780784479346.006 (accessed July 3, 2016).

[24] Y. Yu, M. Wang, D. Zhang, B. Wang, G. Sant, M. Bauchy, Stretched Exponential Relaxation of Glasses at Low Temperature, Physical Review Letters. 115 (2015). https://doi.org/10.1103/PhysRevLett.115.165901.

[25] G.T. Barkema, N. Mousseau, Event-Based Relaxation of Continuous Disordered Systems, Physical Review Letters. 77 (1996) 4358–4361. https://doi.org/10.1103/PhysRevLett.77.4358.

[26] Y. Yu, M. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Bauchy, Thermometer Effect: Origin of the Mixed Alkali Effect in Glass Relaxation, Phys. Rev. Lett. 119 (2017) 095501. https://doi.org/10.1103/PhysRevLett.119.095501.

[27] T. Boutreux, P.G. de Geennes, Compaction of granular mixtures: a free volume model, Physica A: Statistical Mechanics and Its Applications. 244 (1997) 59–67. https://doi.org/10.1016/S0378-4371(97)00236-7.

[28] E. Masoero, E.D. Gado, R. J.-M. Pellenq, S. Yip, F.-J. Ulm, Nano-scale mechanics of colloidal C–S–H gels, Soft Matter. 10 (2014) 491–499. https://doi.org/10.1039/C3SM51815A.

[29] H. Liu, S. Dong, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Effects of Polydispersity and Disorder on the Mechanical Properties of Hydrated Silicate Gels, (n.d.).

[30] H. Manzano, E. Masoero, I. Lopez-Arbeloa, H. M. Jennings, Shear deformations in calcium silicate hydrates, Soft Matter. 9 (2013) 7333–7341. https://doi.org/10.1039/C3SM50442E.

[31] H. Manzano, S. Moeini, F. Marinelli, A.C.T. van Duin, F.-J. Ulm, R.J.-M. Pellenq, Confined Water Dissociation in Microporous Defective Silicates: Mechanism, Dipole Distribution, and Impact on Substrate Properties, J. Am. Chem. Soc. 134 (2012) 2208–2215. https://doi.org/10.1021/ja209152n.

[32] H. Liu, L. Tang, N.M.A. Krishnan, G. Sant, B. Mathieu, Structural Percolation Controls the Precipitation Kinetics of Colloidal Calcium–Silicate–Hydrate Gels, Physical Review Materials. (n.d.).

[33] D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, Elsevier, 2001.

[34] K. Ioannidou, M. Kanduč, L. Li, D. Frenkel, J. Dobnikar, E. Del Gado, The crucial effect of early-stage gelation on the mechanical properties of cement hydrates, Nature Communications. 7 (2016) 12106. https://doi.org/10.1038/ncomms12106.

[35] K. Ioannidou, R. J.-M. Pellenq, E.D. Gado, Controlling local packing and growth in calcium–silicate–hydrate gels, Soft Matter. 10 (2014) 1121–1133. https://doi.org/10.1039/C3SM52232F.

[36] M. Hermes, M. Dijkstra, Jamming of polydisperse hard spheres: The effect of kinetic arrest, EPL. 89 (2010) 38005. https://doi.org/10.1209/0295-5075/89/38005.

[37] A. Donev, I. Cisse, D. Sachs, E.A. Variano, F.H. Stillinger, R. Connelly, S. Torquato, P.M. Chaikin, Improving the Density of Jammed Disordered Packings Using Ellipsoids, Science. 303 (2004) 990–993. https://doi.org/10.1126/science.1093010.

[38] G.D. Scott, D.M. Kilgour, The density of random close packing of spheres, J. Phys. D: Appl. Phys. 2 (1969) 863. https://doi.org/10.1088/0022-3727/2/6/311.

[39] P. Richard, M. Nicodemi, R. Delannay, P. Ribière, D. Bideau, Slow relaxation and compaction of granular systems, Nature Materials. 4 (2005) 121–128. https://doi.org/10.1038/nmat1300.

[40] M. Utz, P.G. Debenedetti, F.H. Stillinger, Atomistic Simulation of Aging and Rejuvenation in Glasses, Physical Review Letters. 84 (2000) 1471–1474. https://doi.org/10.1103/PhysRevLett.84.1471.

[41] G.G. Naumis, Energy landscape and rigidity, Phys. Rev. E. 71 (2005) 026114. https://doi.org/10.1103/PhysRevE.71.026114.

[42] M. Vandamme, F.-J. Ulm, Nanoindentation investigation of creep properties of calcium silicate hydrates, Cement and Concrete Research. 52 (2013) 38–52. https://doi.org/10.1016/j.cemconres.2013.05.006.

[43] M. Davis, N. Thompson, Creep in a Precipitation-Hardened Alloy, Proc. Phys. Soc. B. 63 (1950) 847. https://doi.org/10.1088/0370-1301/63/11/303.

[44] T.W. Lambe, R.V. Whitman, Soil mechanics. Massachusetts institute of technology, John Wiley and Sons, New York, 1969.

[45] E. Masoero, M. Bauchy, E. Del Gado, H. Manzano, R.M. Pellenq, F.-J. Ulm, S. Yip, Kinetic Simulations of Cement Creep: Mechanisms from Shear Deformations of Glasses, in: CONCREEP 10, American Society of Civil Engineers, Vienna, Austria, 2015: pp. 555–564. https://doi.org/10.1061/9780784479346.068.

[46] V. Viasnoff, F. Lequeux, Rejuvenation and Overaging in a Colloidal Glass under Shear, Phys. Rev. Lett. 89 (2002) 065701. https://doi.org/10.1103/PhysRevLett.89.065701.

[47] P. Klug, F. Wittmann, Activation energy of creep of hardened cement paste, Matériaux et Constructions. 2 (1969) 11–16. https://doi.org/10.1007/BF02473650.

Section B. Data-Driven Machine Learning: Make the Data Informative

Chapter 5. Predicting the Dissolution Kinetics of Silicate Glasses by Topology-informed Machine Learning

5.1 Introduction

Machine learning (ML)—a subfield of artificial intelligence—offers a promising route to predict the properties of silicate glasses as a function of their composition [1–7]. Indeed, by "learning" from existing dataset, ML algorithm can infer some complex patterns within the data that would otherwise remain hidden to human eyes [8–10]. As such, ML has previously been used with great success to predict the compositional dependence of the liquidus temperature [1], solubility [2], glass transition temperature [3], stiffness [4], and dissolution kinetics [5] of oxide glasses.

However, data-driven models present several limitations and challenges. (i) The use of ML requires the existence of large, accurate, and consistent datasets (wherein a consistent dataset should comprise data that are measured by the same operation, including the same equipment, operator, protocol, data processing scheme, and environmental conditions), which are not always available [8,11]. (ii) Data-driven models are usually good at "interpolating" data, but typically fail to "extrapolate" data far from the training set [5,10,12]. This is a serious issue as it implies that ML cannot reliably be used to investigate presently unexplored compositional domains that are not explicitly considered during the training phase. This limits the potential of ML for the discovery of novel glasses with significantly improved properties. (iii) Data-driven models do not embed any mechanistic knowledge and, as such, can violate physical laws [8,12]. (iv) Finally, ML-based models are usually complex and hardly interpretable (i.e., they act as "black boxes"). Hence, they usually do not offer any new physical insights [3,5,8]. These issues are challenging to mitigate within traditional machine learning frameworks—wherein traditional descriptors (e.g., glass

composition, interatomic bond energy, etc.) ignore underlying physical and chemical mechanisms and may not properly exhibit a simple and direct relationship with the predicted properties. More generally, when the linkages between the descriptors and the mechanism governing the target property of interest is unclear, the causality of the learned descriptor-property relation is uncertain [13].

Here, to address the challenges facing traditional "*blind machine learning*" (i.e., which does not embed any topological information), we introduce a "*topology-informed machine learning*" paradigm—wherein some features of the network topology are used as descriptors—and apply it to predict the stage I dissolution kinetics (i.e., forward rate, far from saturation) of sodium aluminosilicate glasses [14–16]. Indeed, no universal physics-based model is presently available to predict the dissolution kinetics of silicate glasses (even in stage I). This arises from (i) a lack of knowledge regarding the rate-controlling mechanism of dissolution [14,17–19], (ii) the large number of potential intrinsic (e.g., glass composition) and extrinsic (e.g., temperature, pH, etc.) parameters [5,14,20], and (iii) an incomplete knowledge of the complex, disordered structure of silicate glasses [21–25]. In the present contribution, we show that, by embedding some degree of physics and chemistry, our approach yields a predictive model that is simple (linear), accurate, and transferable to untrained glass compositions.

5.2 Methods

5.2.1 Experimental dissolution rate data

For each glass composition and pH, the dissolution tests conducted by Hamilton *et al.* were carried out on glass powders comprising grain sizes ranging from 74-to-149 μm. These experiments were conducted under static conditions at a surface area to solution volume ratio

(SA/V) of approximately 1.4 to 2.0 cm^{-1} [26]. For each pH, the extent of dissolution was assessed from the concentration of leached SiO_2 (as measured by ICP-AES and ICP-MS) in solution at 5-to-7 regular intervals (for example, 24, 49, 96, 168, and 336 h) of solvent contact. In each case, the pH was recorded before any dissolution and at the time of the dissolution measurement. All the experiments were conducted at 25 °C and ambient pressure. The experimental data present an uncertainty of ±1.5% of the logarithm of the dissolution rates—as estimated from the standard deviation of the dissolution rate data associated with different measurements conducted on the same glass and at constant pH. More details about the measurements can be found in Ref. [26].

5.2.2 Machine learning method

The data points from the training set are first divided into a training and test set (which comprises 30% of the data points). The test set is created by randomly selecting some data points within the training set, while ensuring that the data from the test set are truly unknown to the training set (that is, the pH/compositions combinations used in the test set are not present in the training set). Polynomial regression is then used as a regression method to infer the relationship between inputs and output [9,10]. The least square optimization method is used during the training process of the regression models. We then adopt the 10-fold cross-validation technique [9,10] to optimize the complexity of the model, that is, to identify the maximum polynomial degree of the model. This is accomplished by dividing the initial training set into 10 folds, training the model based on 9 of the folds, and using the remaining fold for validation. This procedure is then repeated 10 times until each of the folds has been used as a validation set. The accuracy of the model (for a given maximum polynomial degree) is then determined by averaging the accuracy of the prediction over all the 10 validation folds. The accuracy of the final model (i.e., with optimal complexity) is

then assessed by computing the relative root square mean square error (RRMSE) by comparing the measured and predicted dissolution rate values DR_i present in the test set [27]:

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\mathrm{DR}_i^{\mathrm{predicted}} - \mathrm{DR}_i^{\mathrm{measured}}\right)^2}{n}} \Bigg/ \left|\frac{\sum_{i=1}^{n}\mathrm{DR}_i^{\mathrm{measured}}}{n}\right| \qquad \textbf{Eq. (5-1)}$$

The intrinsic uncertainty of the dissolution data is here directly embedded within the machine learning framework by incorporating in the training set all the dissolution data obtained for the same glass composition and solution pH (rather than only their average value). This imposes a lower bound of RRMSE = 1.5%, which corresponds to the intrinsic degree of uncertainty of the DR dataset measured in experiments.

5.2.3 Topological constraints enumeration

Topological constraint theory (TCT) describes the disordered network of glasses as a mechanical truss wherein the atoms are connected to each other via some constraints [21,28,29]. TCT considers two kinds of constraints, namely, (i) the radial bond-stretching (BS) constraints that keep the interatomic bond length fixed around their average values and (ii) the angular bond-bending (BB) constraints that fix the average values of the interatomic angles. A previous study recently suggested that the dissolution rate is related to the number of constraints per atom in the "skeleton" network (that is, that formed by the network-forming species, i.e., Si and O here) rather than to the number of constraints per atom in the whole network (that is, including the network-modifying species, i.e., Na here) [30]. Based on this, we enumerate the number of constraints per atom in $(Na_2O)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ as follows. (i) Each Si creates 4 BS constraints with its 4 surrounding O neighbors and 5 BB constraints (i.e., the number of independent angles that needs to be fixed to define the tetrahedral angular environment of Si atoms). Note that, here, the BS

constraints are fully attributed to the cations—so that we do not attribute any BS constraint to the O atoms. (ii) Each tetrahedral Al creates 4 BS constraints with its 4 oxygen neighbors. However, based on previous findings [31], Al atoms do not create any BB constraints—in agreement with the fact that the angular environment of Al atoms is not as well defined as that of Si atoms [32]. (iii) Bridging oxygen (BO) atoms (i.e., surrounded by 2 network-forming cations) form 1 BB constraint. The number of constraints per atom n_c is then calculated by summing the number of constraints created by each element and dividing by the total number of atoms in the skeleton network, namely, Si, Al, BO, NBO (non-bridging oxygen atoms), but excluding Na. The constraints enumeration is summarized in Tab. 5-1. It follows that:

$$n_c = \frac{11-12x+2y}{3-2x+2y} \qquad \textbf{Eq. (5-2)}$$

This metric (n_c) is used as an input (in lieu of x and y) in Model IV.

Similarly, the number of bond-stretching constraints per atom BS is calculated by summing all bond stretching constraints created by each element and dividing by the total number of atoms in the skeleton network:

$$\mathrm{BS} = \frac{4-4x+4y}{3-2x+2y} \qquad \textbf{Eq. (5-3)}$$

The number of bond-bending constraints per atom BB is calculated by summing all bond bending constraints created by each element and dividing by the total number of atoms in the skeleton network:

$$\mathrm{BB} = \frac{7-8x-2y}{3-2x+2y} \qquad \textbf{Eq. (5-4)}$$

The silicon-dominated constraints per atom n_c^{Si} is calculated by summing the number of constraints created by silicon atoms and dividing by the total number of atoms in the skeleton network:

$$n_c^{Si} = \frac{9-9x-9y}{3-2x+2y} \quad \textbf{Eq. (5-5)}$$

The aluminum-dominated constraints per atom n_c^{Al} is calculated by summing the number of constraints created by aluminum atoms and dividing by the total number of atoms in the skeleton network:

$$n_c^{Al} = \frac{8y}{3-2x+2y} \quad \textbf{Eq. (5-6)}$$

In all cases, each input X (i.e., BS, BB, n_c^{Si}, and n_c^{Al}) is transformed into a normalized variable $0 < X' < 1$ as:

$$X' = \frac{X-X_{min}}{X_{max}-X_{min}} \quad \textbf{Eq. (5-7)}$$

where X_{min} and X_{max} are the minimum and maximum values of X, respectively.

Table 5-1. Table summarizing the fraction, coordination number (CN), number of bond-stretching (BS), and number of bond-bending (BB) constraints created by each atomic species in $(Na_2O)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ glasses. Note that $y \geqslant x$ in all glasses, so that all the Al atoms are assumed to be in tetrahedral configuration [33].

Atom	Fraction	CN	BS	BB	BS+BB
Si	$1 - x - y$	4	4	5	9
Al	$2y$	4	4	0	4
Na	$2x$	-	-	-	-
O	$2 - x + y$	-	-	-	-
NBO	$2x - 2y$	1	-	0	0
BO	$2 - 3x + 3y$	2	-	1	1

5.3 Results

5.3.1 Nature of the dataset

To establish our conclusions, we rely on the database developed by Hamilton *et al.* [24,26,34,35], which comprises the forward dissolution rate of a series of aluminosilicate glasses with varying compositions under varying pH conditions. In details, the database comprises dissolution data for two families of glasses, namely, (i) the "Glasses A" series $(Na_2O)_{25}(Al_2O_3)_y(SiO_2)_{75-y}$, with y = 5%, 10%, 15%, 20%, and 25% and (ii) the "Glasses B" series $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$, with x = 12.5%, 16.7%, and 25%. As such, the glass compositions cover both the tectosilicate and peralkaline domains, with varying fractions of non-bridging oxygen atoms. The dissolution kinetics of these glasses is systematically studied in unsaturated aqueous solutions over a wide domain of pH, ranging from pH 1 to pH 12. The dissolution rate is here quantified in terms of the SiO_2 leaching rate (expressed in units of mol/cm^2/s). In total, the database comprises 200 data points [26]. More details can be found in the Methods section. Note that simple metrics (e.g., the fraction of non-bridging oxygen atoms) do not offer any good correlation with the dissolution rate (see Ref. [5]). In particular, all the glasses from the series B are fully charge-compensated and, hence, present a theoretical zero fraction of non-bridging oxygen atoms and yet exhibit varying dissolution rates. This justifies the use of more complex descriptors as presented in the following.

5.3.2 Blind machine learning

We first assess the ability of "blind machine learning" [8,10,12] (that is, which does not embed any physics/chemistry about the dissolution process) to offer realistic prediction of the dissolution kinetics of the aluminosilicate glasses considered herein. To this end, we first consider

as inputs the glass composition (i.e., the molar fractions of Na_2O and Al_2O_3) and the solution pH, while the dissolution rate (DR) is used as output. We then adopt the polynomial regression technique to infer the relationship between inputs and output [9,10]. Indeed, although our previous work on the same DR dataset has shown that more complex machine learning algorithms (e.g., artificial neural network) offer improved performance [5], such complex algorithms do not yield any analytical, easily usable function relating the inputs and output of the model and are poorly interpretable. In contrast, the polynomial regression method eventually yields an analytical model expressing the dissolution rate as a polynomial function of the inputs:

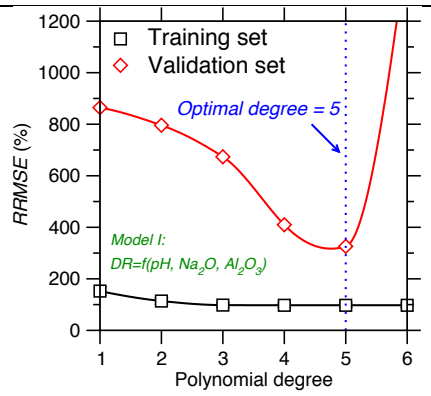
$$\text{Model I:} \qquad \mathrm{DR} = f(\mathrm{pH}, \mathrm{Na_2O}, \mathrm{Al_2O_3}) \qquad \textbf{Eq. (5-8)}$$

In the following, we refer to this model as "Model I." To avoid any overfitting, we divide the database into (i) a training set, which is used to train the model, (ii) a validation set (10% of the data points of the database generated by the cross-validation method [9,10]), which is used to validate the performance of the model and identify the optimal polynomial degree, and (iii) a test set, that is, some data that are kept fully invisible to the model and that are used to assess its ability to predict unknown data. The test is here chosen by randomly selecting 30% of the data points from the database. The accuracy of the prediction is assessed by calculating the relative-root-mean-square-error (RRMSE [27], see Methods section). More details about the machine learning methodology can be found in the Methods section.

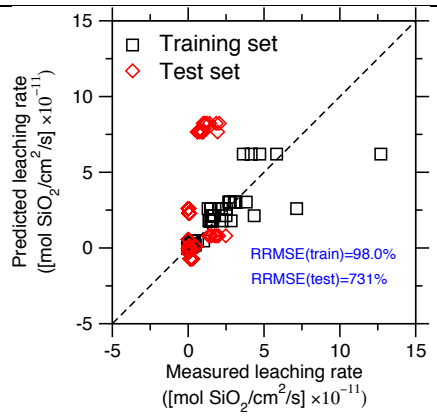
Figure 5-1(a) shows the RRMSE of the training and validation sets as a function of the maximum polynomial degree (p) of the model. We observe that the RRMSE of the training set decreases monotonically upon increasing polynomial degree (i.e., increasing model complexity) and eventually plateaus. This signals that, as the model becomes more complex, it can better interpolate the training set. In contrast, we observe an increase in the RRMSE of the training set

when the polynomial degree is lower than 3, which indicates that these models are too simple to properly describe the relationship between inputs and output. On the other hand, we observe that the RRMSE of the validation set initially decreases with increasing polynomial degree, shows a minimum for degree 5, and eventually increases with increasing degree. This demonstrates that the models incorporating some polynomial terms that are strictly larger than 5 are overfitted. This arises from the fact that, in the case of high degrees, the model starts to fit the "noise" of the training set rather than the "true" overall trend. These results exemplify how the evolution of the RRMSE as a function of the polynomial degree of the model allow us to identify (i) the minimum level of model complexity that is required to avoid underfitting and (ii) the maximum level of model complexity before overfitting. Overall, the optimal polynomial degree (here found to be 5) manifests itself by a minimum in the RRMSE of the validation set.

Figure 5-1(b) shows the dissolution rate values predicted by this model with $p = 5$ for the training and test sets. Overall, we find that blind polynomial regression (Model I) does not accurately capture the relationship between glass composition, pH, and dissolution rate. The RRMSE of the training set is found to be very high (98%), which indicates that the model does not properly interpolate the data used during its training. In turn, the RRMSE of the test set (731%) highlights the fact that this model is largely unable to properly predict the dissolution rate of glasses/pH for which it has not explicitly been trained for. This likely arises from the fact that the relationship between inputs and output is here largely non-linear and, hence, cannot be properly captured by a linear model—in agreement with our previous findings [5]. Note that, considering the low performance of the present model, no effort is here made to understand why the dissolution rates of certain glasses are well predicted, whereas others are not.



(a)



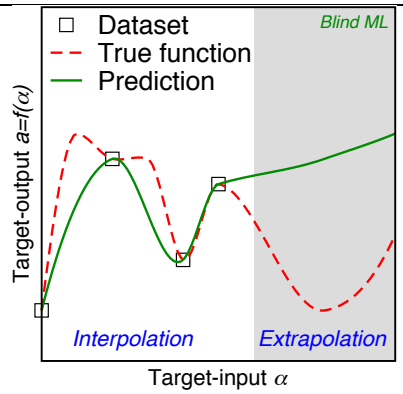
(b)

Figure 5-1: Predictions from "blind" machine learning ("Model I"). **(a)** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates that $p = 5$ is the optimal polynomial degree. **(b)** Predicted dissolution rate for $p = 5$ as a function of the measured dissolution rate.

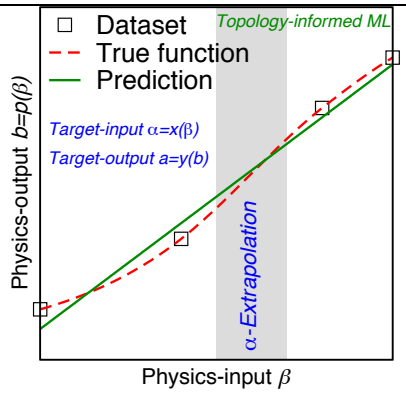
5.3.3 Strategy for topology-informed machine learning

Figure 5-2 illustrates the main idea of "topology-informed" machine learning and how it compares to traditional "blind" machine learning. By being blind to the nature of the mechanism governing the property of interest, traditional blind machine learning ignores (i) which descriptors govern the output property and (ii) the analytical form of the input-output relationship. As illustrated in Fig. 5-2(a), a poor choice of descriptors can result in a complex, highly nonlinear function. Although complex regression algorithms can properly interpolate such nonlinear datasets, they are unlikely to offer realistic predictions extrapolated far from the training set. In contrast, topology-informed ML models are expected to address these limitations by: (i) reducing the dimensionality of the problem (since several glasses with varying compositions can present the

same topology and, hence, similar dissolution kinetics), (ii) simplifying the trained models (since the number of descriptors is decreased), and (iii) linearizing the relationship between inputs and output. As illustrated in Fig. 5-2(b), relying on a topological fingerprint (rather than traditional descriptors) is expected to facilitate extrapolations far from the training set.



(a)



(b)

Figure 5-2: Schematic illustrating the ability (or inability) to extrapolate predictions far from the training set of **(a)** traditional blind machine learning (trained based on arbitrary descriptors α) and **(b)** topology-informed machine learning (trained based on topological descriptors β). In both panels, the dashed red curve represents the true function relating the inputs to the targeted output. The squares indicate the known points from the training set. The solid green curve represents the "guessed" function interpolated by the ML model. The grey window indicates a range of systems (i.e., specific values of descriptors α) that is not represented within the training set and for the predictions from the ML models are tested. Note that this window is outside the training set in panel (a), but not in panel (b)—since several systems with different descriptors α may present the same topology.

In detail, to address the intrinsic limitations of blind machine learning highlighted in Fig. 5-1 and 5-2, we adopt the following strategy. (i) First, we focus on the polynomial regression

method since more complex machine learning algorithms (e.g., artificial neural network or random forest [9,10]) offer poor interpretability [8]. Rather, the polynomial regression yields an analytical function, which, in turn, can serve to infer some of the underlying physics of the dissolution mechanism. (ii) Second, we attempt to "linearize" the relationship between inputs and output based our physical understanding of the dissolution process. This is based on the idea that linear models are expected to be more likely to offer a good transferability to unknown inputs and to potentially yield some useful physical insights [8,10,12]. (iii) Third, we attempt to identify some relevant reduced-dimensionality descriptors capturing the effect of the atomic structure of the glass on dissolution rate that can be used as inputs. This is based on the idea that, although the dissolution kinetics of glasses is controlled by their composition (at fixed thermal history) for a given set of environment conditions (T, pH and solution composition [36–39]), the knowledge of the structure of the atomic network makes it possible to decipher the relationship between composition and dissolution rate—so that it should be easier for machine learning algorithms to infer the relationship between "structure and dissolution rate" than between "composition and dissolution rate." In the following, we present how these topology-informed ingredients allow us to derive less complex, yet more accurate predictive models.

5.3.4 Linearization of the inputs/output relationship

In an attempt to linearize the relationship between the inputs and output of the model, we first note that, in general, the dissolution rate is an exponential (rather than linear) function of pH and composition. This can be illustrated from the fact that, based on transition state theory, the Aagaard-Helgeson model expresses the forward dissolution rate in terms of (i) the activity of H^+ ions, which, in turn, is an exponential function of pH [40], and (ii) an Arrhenius term $\exp(-E_a/RT)$,

wherein the activation energy has recently be suggested to be a function of the number of topological constraints per atom in the network, which, in turn, is often a linear function of composition [21,36,37]. Based on this fact, it follows that one can increase the degree of linearity of the relationship between inputs and output by predicting the logarithm of the dissolution rather than the dissolution rate itself (referred to as "Model II" thereafter):

$$\text{Model II:} \qquad \log(\text{DR}) = f(\text{pH}, \text{Na}_2\text{O}, \text{Al}_2\text{O}_3) \qquad \textbf{Eq. (5-9)}$$

We find that, by using Model II, the prediction accuracy is significantly improved when the polynomial degree *p* decreases to 3. To further enhance the degree of linearity of the inputs/output relationship, we now consider the dependence of the dissolution on pH. As illustrated in Fig. 5-3, the dissolution rate exhibits a fairly bilinear V-shape dependence on pH, with a minimum in neutral condition (pH 7) [36,38]. This is an issue as the description of a bilinear function in terms of a sum of polynomials requires the use of high degrees to account for the break in slope. As an alternative route, we define two new input variables, namely, pH_{acid} and pH_{base}, which are defined as $\text{pH}_{\text{acid}} = \max(0; 7 - \text{pH})$ and $\text{pH}_{\text{base}} = \max(0; \text{pH} - 7)$. Note that these inputs contain the same information of the pH variable but allow us to describe the *linear* evolution of the dissolution rate with respect to pH_{acid} and pH_{base} for pH < 7 and pH > 7, respectively, rather than the bilinear evolution of the dissolution with respect to pH (see Fig. 5-3). Note that the variables pH_{acid} and pH_{basic} are equal to 0 for pH > 7 and pH < 7, respectively, so that only one of these variables at a time is non-zero. Model III expresses the logarithm of the dissolution rate in terms of the glass composition and these two new variables:

$$\text{Model III:} \qquad \log(\text{DR}) = f(\text{pH}_{\text{acid}}, \text{pH}_{\text{base}}, \text{Na}_2\text{O}, \text{Al}_2\text{O}_3) \qquad \textbf{Eq. (5-10)}$$

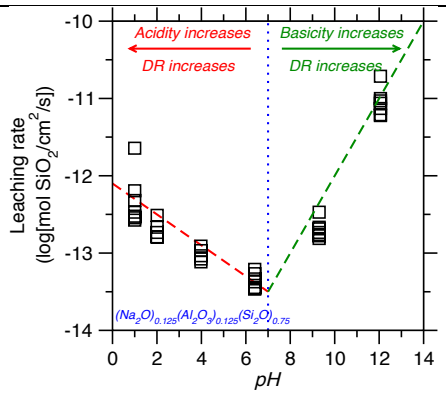
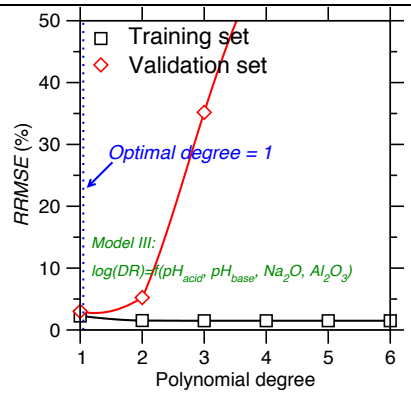
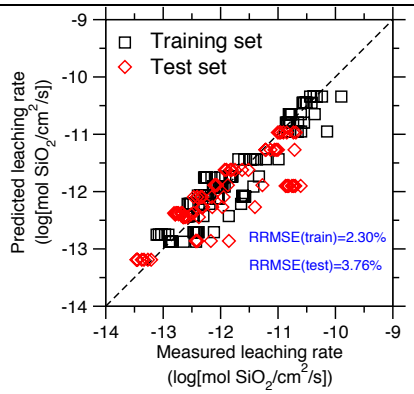


Figure 5-3: Measured dissolution rate of a $(Na_2O)_{0.125}(Al_2O_3)_{0.125}(SiO_2)_{0.75}$ glass as a function of pH [26].



(a)



(b)

Figure 5-4: Predictions from machine learning while explicitly accounting for the exponential dependence of the dissolution rate on the inputs and capturing the distinct acidic and caustic regimes ("Model III"). **(a)** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **(b)** Predicted dissolution rate for $p = 1$ as a function of the measured dissolution rate.

Figure 5-4(a) shows the RRMSE of the training and validation sets as a function of the maximum polynomial degree p for Model III. Importantly, we find that the explicit description of the bilinear dependence of the dissolution rate on pH allows us to further reduce the complexity of the model since the RRMSE of the validation set shows a minimum for $p = 1$. This indicates that Model III can express the dissolution rate through a simple, linear relationship. In addition to decreasing the complexity of the model, Model III also offers an increased degree of accuracy since the RRMSE of the test set is found to be 3.76% (as compared to 731% for Model I, see Fig. 5-4(b)). These results illustrate how the linearization of the relationship between inputs and output based on our physical/chemical understanding of the dissolution process can results in the training of a less complex, yet more accurate model.

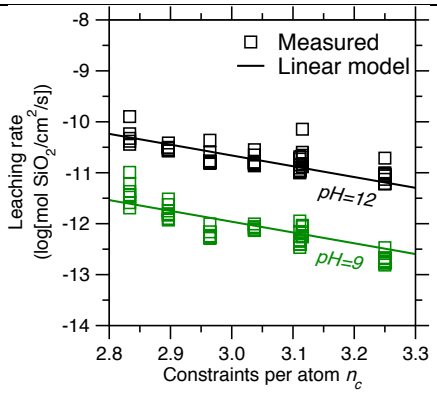


Figure 5-5: Dissolution rate of the silicate glasses considered herein as a function of the number of topological constraints per atom for pH = 9 and 12.

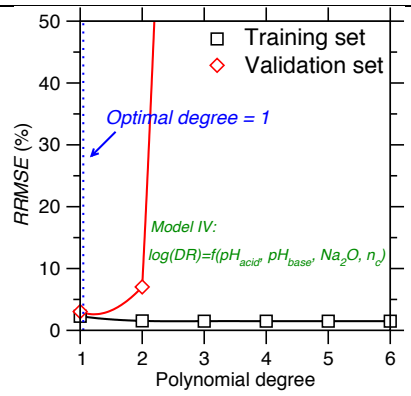
5.3.5 Topology-informed reduced-dimensionality descriptors

We now attempt to further increase the accuracy of the model by identifying a structural "fingerprint" of the structure of the atomic network—which is based on the idea that the structure of the atomic network of a glass has a first order effect on its dissolution kinetics. To this end, we adopt the framework of topological constraint theory (TCT), which describes complex disordered

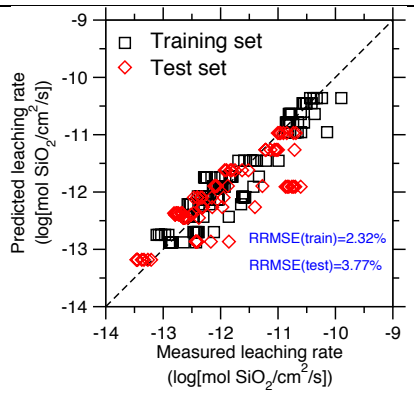
network as mechanical trusses, wherein some nodes (the atoms) are connected to each other by some topological constraints (the chemical bonds) [21,28,29,41]. Based on this framework, the number of topological constraints per atom (n_c) has been shown to offer a useful reduced-dimensionality descriptor that captures the connectivity of the atomic network and, hence, can be used to predict various glass properties, e.g., hardness, stiffness, fracture toughness, glass transition temperature, fragility, etc. [21,42–47]. Importantly, the effective activation energy of dissolution for a fixed pH has recently been suggested to be proportional to n_c [30,31,37,39,48–52]. Based on these findings, we compute the number of topological constraints of the rigid aluminosilicate network n_c for each glass (see Methods section) and use it as a descriptor of the atomic structure. As shown in Fig. 5-5, we observe that, at fixed pH, the dissolution rate is indeed largely correlated to n_c, which supports the use of this metric as an input to the model. We then define Model IV, which expresses the logarithm of the dissolution rate in terms of pH, n_c, and the fraction of network modifiers (i.e., Na_2O)—since the network modifiers are not explicitly accounted for in the number of topological constraints of the rigid aluminosilicate network (see Methods) [30]:

$$\text{Model IV:} \qquad \log(\mathrm{DR}) = f(\mathrm{pH_{acid}}, \mathrm{pH_{base}}, n_c, \mathrm{Na_2O}) \qquad \textbf{Eq. (5-11)}$$

Figure 5-6(a) shows the RRMSE of the training and validation sets as a function of the maximum polynomial degree p for Model IV. Like Model III, we note that a linear model (i.e., $p = 1$) offers the best performance. As shown in Fig. 5-6(b), Model IV is able to (i) properly interpolate the training set and (ii) predict realistic values for the test set. Nevertheless, we note that the overall degree of accuracy remains comparable to that offered by Model III. In particular, select points appear to systematically act as outliners in all the models considered herein and, hence, might be experimental artefacts.



(a)



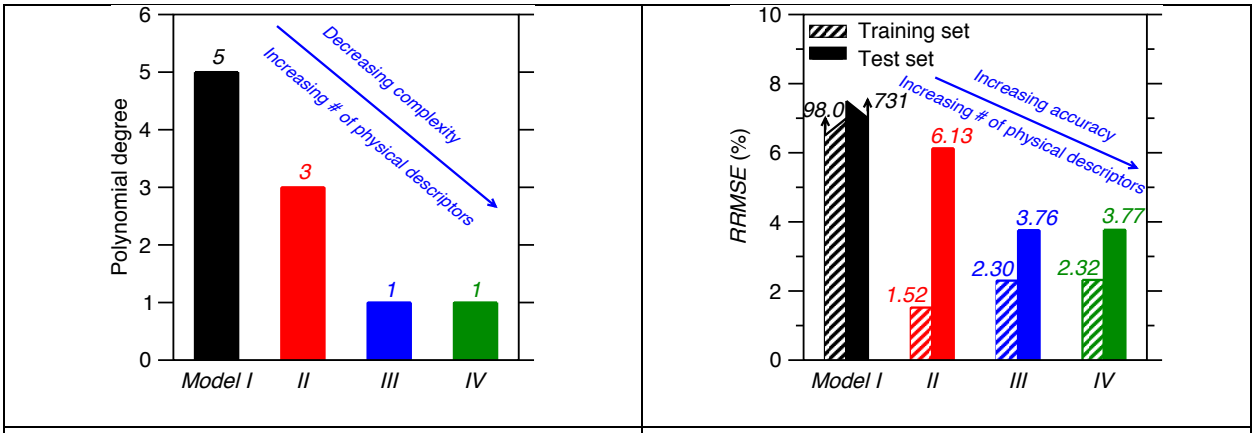
(b)

Figure 5-6: Predictions from "topology-informed" machine learning, that is, by explicitly accounting for the exponential dependence of the dissolution rate on the inputs, capturing the distinct acidic and caustic regimes, and describing the glass structure in terms of the number of topological constraints per atom n_c ("Model IV"). **(a)** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **(b)** Predicted dissolution rate for $p = 1$ as a function of the measured dissolution rate.

5.3.6 Overcoming the tradeoff between accuracy and simplicity in machine learning

ML-based models usually exhibit a balance between accuracy, simplicity, and interpretability (i.e., the degree to which a human can understand the outcome produced by the model) [8–10]. Indeed, simple and interpretable models (e.g., polynomial regression) usually offer limited accuracy, whereas more advanced models (e.g., random forest or artificial neural network) offer increased levels of accuracy but often come with higher complexity and lower interpretability [5,10,53]. In general, simpler and more interpretable models are desirable as (i) simpler models are less likely to overfit small datasets, (ii) simpler models are usually more computationally-

efficient, and (iii) more interpretable models are more likely to offer some new insights into the underlying physics governing the relationship between inputs and outputs. Figure 5-7 shows the complexity (captured by the optimal polynomial degree) and accuracy (captured by the RRMSE) of the different models considered herein. Overall, we find that embedding topological descriptors yields models that are less complex and more accurate. This establishes topology-informed machine learning as a promising route to overcome the tradeoff between accuracy and simplicity, which are otherwise often mutually exclusive [5,10,53].



(a) **(b)**

Figure 5-7: (a) Complexity (as captured by the polynomial degree) and **(b)** accuracy (as captured by the relative root square mean square error, RRMSE) of the "blind" and "topology-informed" machine learning models described herein.

5.4 Discussion

We now discuss the interest of using topology-informed reduced-dimensionality descriptors as inputs to the ML model. As shown in Fig. 5-5, the number of constraints per atom n_c offers a powerful reduced-dimensionality since it allows us to describe the evolution of the dissolution rate in terms of one variable (i.e., n_c) instead of two (that is, the molar fractions of Na_2O and Al_2O_3). However, as shown in Fig. 5-7, we find that Model III (which is blind to the topology

of the atomic network) offers a level of accuracy that is comparable to that offered by Model IV (which embeds n_c as an explicit input). To further understand this point, we now assess whether Model III is able to "learn" by itself that the dissolution rate can be described by the reduced-dimensionality parameter n_c. To this end, we analyze the coefficients of the final linear function yielded by Model III, which relates –log(DR) to the pH and the molar fractions of Na_2O and Al_2O_3. This model can be expressed as:

$$\mathrm{DR} = F(\mathrm{pH}) \exp(a[\mathrm{Na_2O}] + b[\mathrm{Al_2O_3}]) \qquad \textbf{Eq. (5-12)}$$

where F(pH) is a function that depends only on the pH of the solution and a and b are some coefficients of the model. On the other hand, Ref. [37] suggests that the dissolution rate can be expressed as:

$$\mathrm{DR} = \mathrm{DR_0}(\mathrm{pH}) \exp\left[\frac{-n_c E_0}{RT}\right] \qquad \textbf{Eq. (5-13)}$$

where DR_0(pH) is the dissolution rate when $n_c = 0$, E_0 is activation energy needed to break a unit constraint per atom, R is the perfect gas constant, and T is the temperature.

A comparison between Eqs. (5-12) and (5-13)—i.e., by setting equal their respective exponent terms—allows us to determine the number of topological constraints per atom n_c^{guessed} that is "guessed" by Model III as a function of the glass composition. As shown in Fig. 5-8, we find that Model III is able to infer how the number of constraints depends on the glass composition (see Methods section), which explains why Model III and Model IV eventually offer the same level of accuracy. This demonstrates that, in the present case, ML is able to learn by itself some chemical rules governing the number of topological constraints created by each atom. Note that the number of constraints per atom (n_c) depends not only on glass composition, but also on some "chemical knowledge" of the system, that is, (i) the coordination number of each atom, (ii) the energy of each chemical bond, which can be active or thermally-broken, and (iii) the directionality

of each interatomic bond (i.e., covalent vs. ionic), which governs the existence of bond-bending constraints. In that sense, it is notable that the ML model is able to properly "guess" all these chemical features and how they govern the dissolution rate. As discussed below, this is permitted by the fact that, here, the training set homogeneously covers all the range of the possible glass compositions. More generally, these results exemplify how an interpretable ML model can offer some physical insights into the relationship between inputs and output—which would not be possible with a less interpretable model (e.g., artificial neural network).

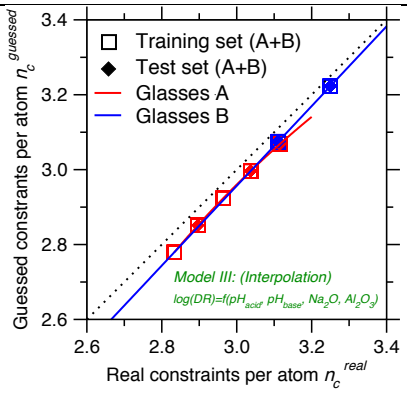


Figure 5-8: Number of topological constraints per atom n_c "guessed" by Model III (which is blind to the topology of the atomic network) as a function of the real value of n_c—wherein the training set randomly covers the whole range of glass composition and solution pH. The red and blue lines indicate the guessed n_c values for the two families of glasses considered herein, namely, $(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$ (Glasses A) and $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$ (Glasses B).

We now assess whether the models considered herein can be used to extrapolate predictions, that is, to predict the dissolution rate of glasses with compositions that are different from those used during the training of the model. To this end, rather than randomly choosing data from database to serve as a test set, we purposely select the data from the Glasses A series as a training set and those from the Glasses B series as a test set. In other words, (i) we train our models based

on the dissolution rate data of the first series of glasses with varying Na/Al molar ratios, namely, $(Na_2O)_{25}(Al_2O_3)_y(SiO_2)_{75-y}$ and (ii) we test the ability of the models to predict the dissolution rate of the second series of fully charge-compensated glasses with varying fractions of Na_2O, namely, $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$. In this scenario, the training set does not homogeneously sample the range of glass composition, which allows us to determine whether the models are able to extrapolate predictions from their training sets. Note that these two families of glasses exhibit significantly different structures, namely, (i) Glasses A exhibit varying degrees of polymerization and present some non-bridging oxygen (NBO) atoms, whereas (ii) Glasses B are fully-compensated and theoretically do not comprise any NBO. In addition, the training set (Glasses A) presents a fixed fraction of Na_2O, so that the test set (Glasses B, with varying fractions of Na_2O) is truly unknown to the model.

Figure 5-9 shows the dissolution rate data predicted by Model III ("topology-blind") and Model IV ("topology-informed") based on the above-mentioned training scenarios. In both cases, the prediction error distribution of the training set is centered around 0 with a standard deviation that is close to experimental uncertainty (i.e., $\pm$0.2 log[mol $SiO_2/cm^2/s$]) (see Fig. 5-9(c)). This indicates that both models are able to properly interpolate the training set (i.e., Glasses A). In contrast, we find that the test set RRMSE of Model IV is lower than that offered by Model III. In addition, we note that the prediction error distribution is around 0 in Model IV, but shows a systematic deviation from 0 in Model III (see Fig. 5-9(c)). This signals that the topology-informed Model IV shows an enhanced ability to extrapolate predictions far from the training set.

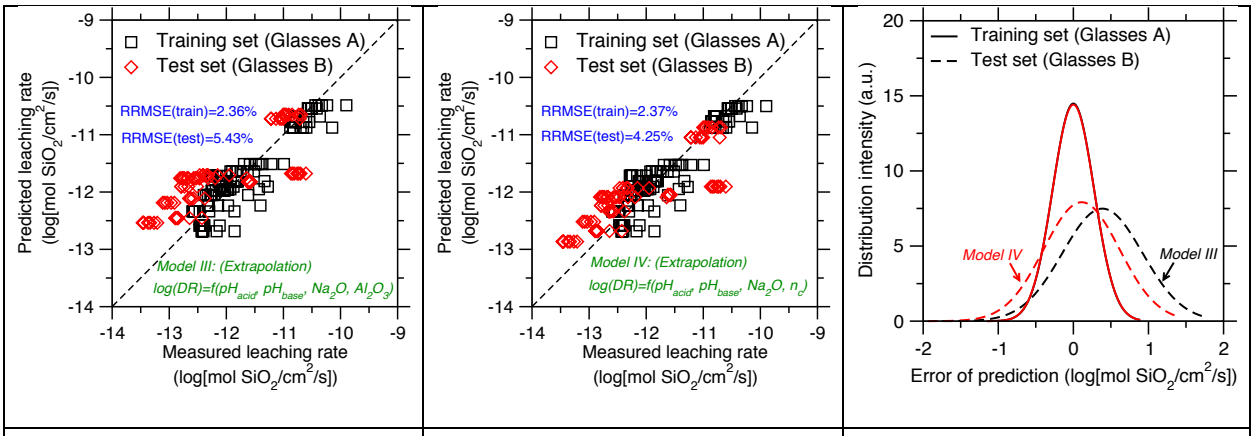


Figure 5-9: Dissolution rate predicted by **(a)** "topology-blind" machine learning (Model III) and **(b)** "topology-informed" machine learning (Model IV) as a function of the measured dissolution rate—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set). **(c)** Distribution of prediction error for the training (solid line) and test sets (dash line) offered by Models III (black) and IV (red), respectively. The error is defined as the difference between predicted and measured dissolution rate.

To further understand how explicitly using the number of constraints per atom n_c as a reduced-dimensionality input enhances the extrapolability of Model IV, we assess whether Model III is still able to "guess" by itself the compositional-dependence of the number of constraints per atom when the training set does not homogeneously sample the range of glass compositions. Figure 5-10 shows the number of constraints per atom "guessed" by Model III. We find that, here, Model III fails to properly infer the compositional evolution of n_c. This arises from the fact that, in this case, the training set does not homogeneously sample the whole domain of glass compositions—so that it is unable to properly capture how the glass composition governs the number of constraints per atom over the entire compositional domain.

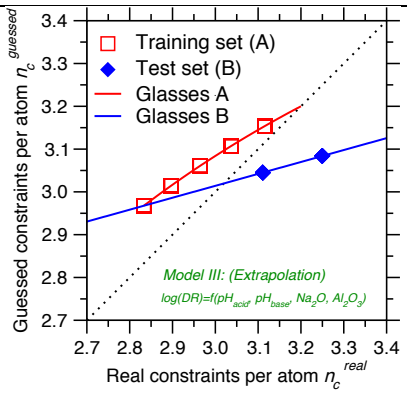


Figure 5-10: Number of topological constraints per atom n_c "guessed" by Model III (which is blind to the topology of the atomic network) as a function of the real value of n_c. The red and blue lines indicate the guessed n_c values for the two families of glasses considered herein, namely, $(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$ (Glasses A) and $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$ (Glasses B), respectively. Here, the dissolution data of Glasses A are used as a training set to predict the dissolution kinetics of Glasses B (test set).

Overall, the fact that training the ML model explicitly based on the number of constraints per atom n_c rather than based on the glass composition enhances the potential for extrapolation can be understood as follows. To offer accurate predictions, topology-blind models (e.g., Model III) have to infer how each elementary oxide (e.g., NaO_2 and Al_2O_3) governs the dissolution rate. This requires the use of a large training set that homogeneously sample all the possible glass compositions. In contrast, topology-informed models (Model IV) only have to infer the relationship between the n_c and the dissolution rate. It follows that, once the relationship between n_c and the dissolution rate is properly parameterized, the model will be able to properly predict the dissolution rate of new unknown glass compositions, provided that their number of constraints n_c is similar to that of some glasses of the training set—based on the idea that two glasses with different composition but similar n_c values will exhibit a comparable dissolution rate. As such,

topology-informed models only need to be trained with a relatively small training set comprising different glasses with varying n_c values to be able to properly predict the dissolution rate of new glasses with compositions that are unknown to the model. This is illustrated by Fig. 5-11, which shows that here, some of the glasses of the B series have a number of constraints per atom n_c that is similar to some of glasses of the A series—so that Model IV (topology-informed) succeeds in predicting their dissolution rate while Model III (topology-blind) does not. This suggests that the use of topological inputs capturing into a single metric (n_c) some details of the glass structure makes it possible to reduce the dimensionality of the problem and, thereby, to train predictive models based on limited datasets.

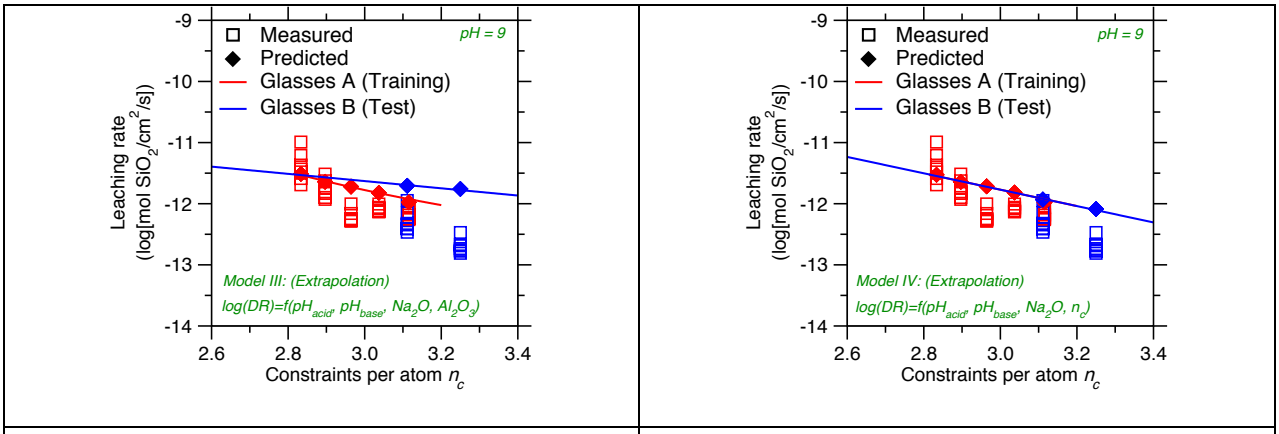


Figure 5-11: Dissolution rate predicted by **(a)** "topology-blind" machine learning (Model III) and **(b)** "topology-informed" machine learning (Model IV) as a function of the number of topological constraints per atom n_c for pH 9—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set). The measured dissolution rates are added for comparison.

We now further assess the degree of transferability of our topology-informed machine learning model by testing its ability to predict the dissolution rate of pure glassy silica (SiO_2, taken from Ref. [37]). It is worth mentioning that, although the composition of this glass may look similar the those of the training set (i.e., Glasses A), pure glassy silica often exhibits unique, anomalous behaviors. For instance, it is notable that the dissolution rate of SiO_2 (or logarithm thereof) cannot be predicted as a linear extrapolation of those of Glasses B $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$ toward $x \rightarrow 0$. As shown in Fig. 5-12, we find that our topology-informed machine learning model offers an excellent prediction of the dissolution rate of glassy silica (with RRMSE = 1.66%). It is notable that, although it is trained for glasses comprising a fixed fraction (25%) of Na_2O, our model is able to accurately predict the dissolution rate of pure silica. These results exemplify how adopting topological descriptors enables extrapolations far from the training set—although it will certainly be desirable in the future to test the predictions of this model to some additional families of silicate glasses (e.g., borosilicate, phosphosilicate, etc.).

Note that traditional machine learning approaches typically rely on a large number of descriptors (e.g., molar masses, bond energy, atomic charges, field strength, etc.), which can be *a posteriori* be filtered out to reduce the complexity of the model (e.g., using LASSO [54]). Although using a large number of descriptors can increase the ability of the model to interpolate complex data, this comes with several challenges, namely, (i) the computational burden required to filter out irrelevant descriptors is increased, (ii) certain descriptors can appear as being insignificant when taken individually, but may become very useful when combined with each other, (iii) models relying on a large number of descriptors typically require large training sets, (iv) a larger number of descriptors usually increase the complexity of the model, (v) a larger number of descriptors usually decrease the interpretability of the model, and (vi) the use of a large number of descriptors

can result in some degree of overfitting, which, in turn, tends to decrease the extrapolability of the model. In contrast, adopting a topological fingerprint of the atomic network filters out some of the structural details. As such, the use of topological descriptors only may not fully capture some of the fine details of the relationship between composition and dissolution kinetics, but, nevertheless, we find here this level of simplification/filtering to be key in enhancing the extrapolability of the trained models.

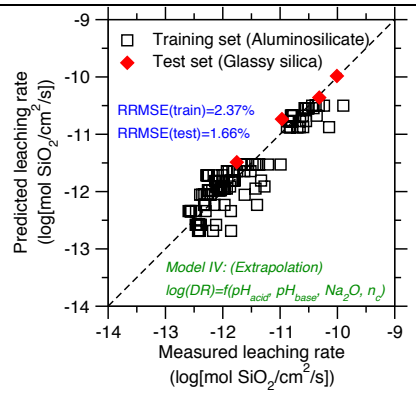
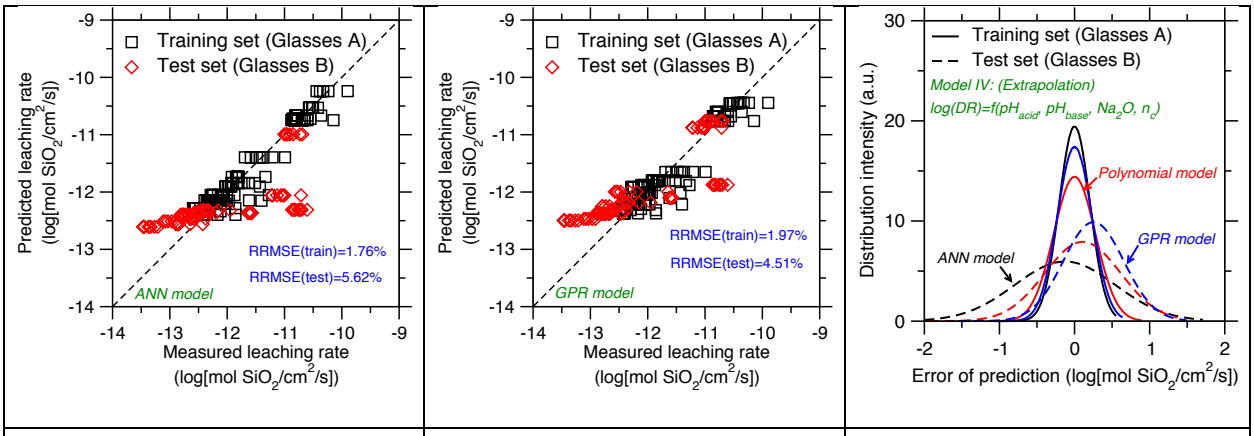


Figure 5-12: Dissolution rate predicted by "topology-informed" machine learning (Model IV) as a function of the measured dissolution rate—wherein the dissolution data of sodium aluminosilicate Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of glassy silica (SiO_2, test set).

Finally, we discuss in terms of (i) model accuracy and (ii) interpretability the choice of using polynomial regression (rather than more complex regression techniques) within the topology-informed machine learning framework presented herein. To this end, we consider the artificial neural network (ANN [55]) and Gaussian process regression (GPR [56]) techniques. The ANN model used herein is a multilayer perceptron model including one hidden layer made up of 20 neurons, while the GPR model used herein is a nonparametric regression model adopting the Matern-type kernel, the noise level of dataset being set as 0.01. Both of these models are trained

with topological descriptors (model IV). We assess their potential for extrapolation by training them based on Glasses A and testing their ability to predict the dissolution rates of Glasses B (see above). As expected, we find that both the ANN and GPR models can very accurately interpolate the training set. In both cases, the RRMSE of the training set is below 2%, which is smaller than that offered by polynomial regression (2.4%). We note that the distribution of the prediction error is centered around 0 and is sharper than that offered by polynomial regression (see Fig. 5-13(c)). This arises from the fact that, as compared to polynomial regression, both the ANN and GPR models exhibit higher complexities, that is, higher numbers of adjustable parameters. This complexity provides them with more flexibility to interpolate fine details of the training set.

However, we find that both the ANN and GPR models do not offer satisfactory predictions for the test set (see Figs. 5-13(a) and 5-13(b)). In detail, the RRMSE of the test set offered by ANN and GPR is 5.62% and 4.51%, respectively, which are both higher than that offered by polynomial regress (i.e., 4.25%, see Fig. 5-9(b)). Notably, a visual inspection of Figs. 5-13(a) and 5-13(b) and the analysis of the distribution of the prediction error (see Fig. 5-13(c)) reveals that both ANN and GPR exhibit a systematic error when predicting the test set—especially for slowly-dissolving glasses, whose dissolution rate tends to be overpredicted. This poor extrapolability can be understood from the fact that both ANN and GPR are intrinsically nonlinear and, hence, do not capture the linear dependence of the logarithm of the dissolution rate on the number of constraints per atom. Such nonlinearity can clearly be observed in Figs. 5-13(a) and 5-13(b). In contrast, polynomial regression intrinsically relies on a linear formulation and, as such, offers more realistic predictions far from the training set. These results exemplify that, in addition of informing the choice of the descriptors, our physical understanding of the underlying mechanism can also guide the choice of the regression technique.



(a) **(b)** **(c)**

Figure 5-13: Dissolution rate predicted by "topology-informed" machine learning (Model IV) using **(a)** Artificial Neural Network (ANN) and **(b)** Gaussian Process Regression (GPR) as a function of the measured dissolution rate—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set). **(c)** Distribution of the prediction error for the training (solid line) and test set (dash line) by using the ANN (black) and GPR models (blue), respectively. The results offered by polynomial regression are added for reference. The error is here defined as the difference between predicted and measured dissolution rates.

As a notable advantage over more complex regression techniques, polynomial regression offers a high degree of interpretability, which, in turn, can offer some physical insights into the nature of the relationship between inputs and outputs. To illustrate this point, we further expand the number of topological descriptors and use our ML model to assess their weight in governing the dissolution kinetics. To this end, we construct two new "topology-informed" models (referred to as Model IV-a and IV-b) by decomposing the term "constraints per atom (n_c)" into several contributions:

Model IV-a: $\log(\mathrm{DR}) = f(\mathrm{pH_{acid}}, \mathrm{pH_{base}}, \mathrm{BS}, \mathrm{BB})$ **Eq. (5-14)**

Model IV-b: $\log(\mathrm{DR}) = f(\mathrm{pH_{acid}}, \mathrm{pH_{base}}, n_c^{\mathrm{Si}}, n_c^{\mathrm{Al}})$ **Eq. (5-15)**

In detail, Model IV-a investigates the relative weights of the radial bond-stretching (BS) and angular bond-bending (BB) constraints, whereas Model IV-b investigates the relative weights of the constraints created by Si and Al atoms (n_c^{Si} and n_c^{Al}, respectively). Note n_c = BS + BB (see Methods section), so that the original Model IV assumes that radial and angular constraints have the same weight, and so do the constraints created by different elements.

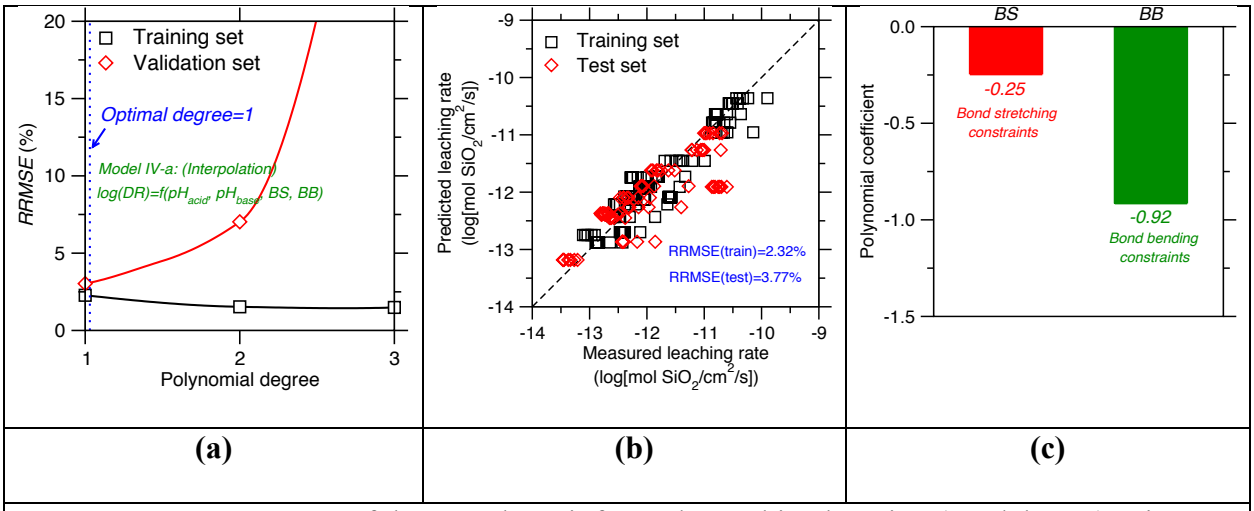


Figure 5-14: Outcomes of the "topology-informed" machine learning (Model IV-a) using as inputs the numbers of bond stretching constraints per atom (BS) and bond bending constraints per atom (BB). **(a)** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree *p*. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **(b)** Predicted dissolution rate (for $p = 1$) as a function of the measured dissolution rate. **(c)** Coefficients of the polynomial model associated with the BS and BB inputs. Note that the BS and BB input values are normalized in the training process to ensure that the model coefficients reflect the contribution of each input to the dissolution rate.

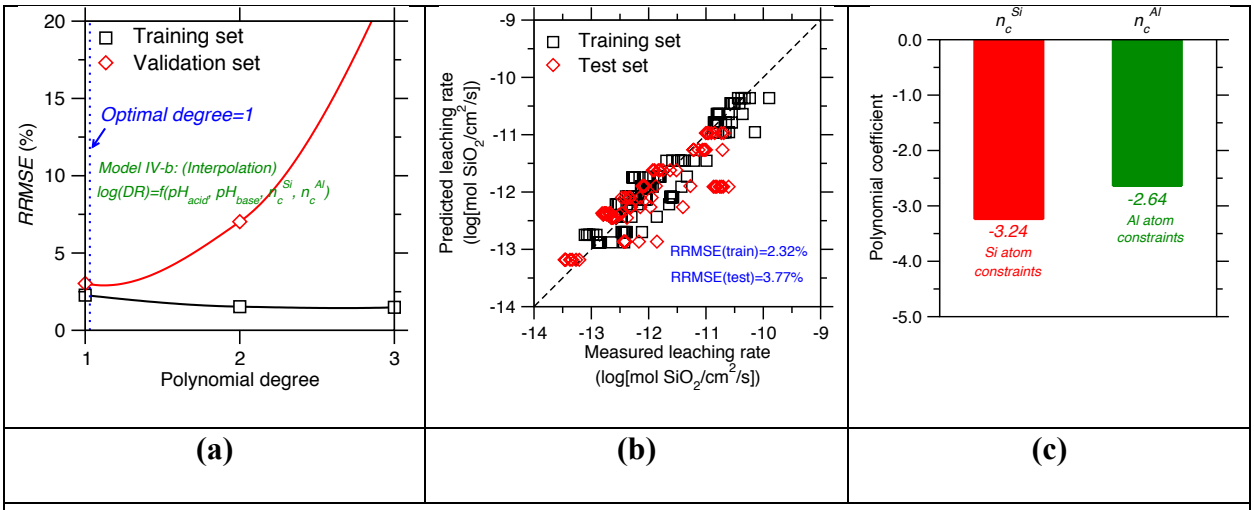


Figure 5-15: Outcomes of the "topology-informed" machine learning (Model IV-b) using as inputs the number of constraints per atom created by silicon (n_c^{Si}) and aluminum (n_c^{Al}). **(a)** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **(b)** Predicted dissolution rate (for $p = 1$) as a function of the measured dissolution rate. **(c)** Coefficients of the polynomial model associated with the n_c^{Si} and n_c^{Al} inputs. Note that, the n_c^{Si} and n_c^{Al} input values are normalized in the training process to ensure that the model coefficients reflect the contribution of each input to the dissolution rate.

Figures 5-14 and 5-15 show the outcomes of Models IV-a and IV-b. First, we find that both models present an optimal degree of 1 (see Figs. 5-14(a) and 5-15(a)). This highlights that the relationship between network topology and the logarithm of the dissolution rate is intrinsically linear. Second, we observe that both models properly interpolate the dataset, with a level of accuracy that is comparable to that offered by the original Model IV (see Figs. 5-14(b) and 5-15(b)). The coefficients of the polynomial regression models can then be interpreted as the weight of each type of constraints in governing the dissolution kinetics. We first note that all the

coefficients are negative (see Figs. 5-14(c) and 5-15(c)), which confirms that all the topological constraints, whatever their nature, tend to decrease the dissolution rate. Interestingly, we find that the angular bond-bending BB constraints present a larger weight than the linear bond-stretching BS constraints (see Figs. 5-14(c)). This finding is confirmed by the fact that the topological constraints created by Si atoms have a larger weight than those created by Al atoms (see Figs. 5-15(c))—since Al atoms do not create any angular constraints (see Methods section) [32]. Overall, these results signal that bond-bending constraints have more influence than radial ones on the dissolution kinetics. This suggests that the dissolution kinetics is strongly affected by the directionality of the interatomic bonds. We note that insights of this nature would be challenging to obtain from more complex, less interpretable "black-box" machine learning models (e.g., ANN).

5.5 Conclusions

Overall, these results show that embedding some physical and chemical descriptors within ML models can increase the degree of linearity of the input/output relationship and reduce the dimensionality of the model. This establishes topology-informed machine learning as a promising route to address some of the limitations of traditional blind machine learning, namely, by (i) reducing the complexity and increasing the interpretability of the trained models, (ii) limiting the need for large training sets, and (iii) enhancing the ability of the models to extrapolate predictions far from their training sets.

5.6 References

[1] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Accelerating the Design of Functional Glasses through Modeling, Chem. Mater. 28 (2016) 4267–4277. https://doi.org/10.1021/acs.chemmater.6b01054.

[2] D.S. Brauer, C. Rüssel, J. Kraft, Solubility of glasses in the system P2O5–CaO–MgO–Na2O–TiO2: Experimental and modeling using artificial neural networks, Journal of Non-Crystalline Solids. 353 (2007) 263–270. https://doi.org/10.1016/j.jnoncrysol.2006.12.005.

[3] D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. Zanotto, Predicting glass transition temperatures using neural networks, Acta Materialia. 159 (2018) 249–256. https://doi.org/10.1016/j.actamat.2018.08.022.

[4] K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, M. Bauchy, Prediction of Silicate Glasses' Stiffness by High-Throughput Molecular Dynamics Simulations and Machine Learning, ArXiv:1901.09323 [Cond-Mat, Physics:Physics]. (2019). http://arxiv.org/abs/1901.09323 (accessed February 12, 2019).

[5] N.M.A. Krishnan, S. Mangalathu, M.M. Smedskjaer, A. Tandia, H. Burton, M. Bauchy, Predicting the dissolution kinetics of silicate glasses using machine learning, Journal of Non-Crystalline Solids. 487 (2018) 37–45. https://doi.org/10.1016/j.jnoncrysol.2018.02.023.

[6] M.C. Onbaşlı, A. Tandia, J.C. Mauro, Mechanical and Compositional Design of High-Strength Corning Gorilla® Glass, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling: Applications: Current and Emerging Materials, Springer International Publishing, Cham, 2018: pp. 1–23. https://doi.org/10.1007/978-3-319-50257-1_100-1.

[7] E.D. Cubuk, R.J.S. Ivancic, S.S. Schoenholz, D.J. Strickland, A. Basu, Z.S. Davidson, J. Fontaine, J.L. Hor, Y.-R. Huang, Y. Jiang, N.C. Keim, K.D. Koshigan, J.A. Lefever, T. Liu, X.-G. Ma, D.J. Magagnosc, E. Morrow, C.P. Ortiz, J.M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K.N. Nordstrom, P.E. Arratia, R.W. Carpick, D.J. Durian, Z. Fakhraai, D.J. Jerolmack, D. Lee, J. Li, R. Riggleman, K.T. Turner, A.G. Yodh, D.S. Gianola, A.J. Liu, Structure-property relationships from universal signatures of plasticity in disordered solids, Science. 358 (2017) 1033–1037. https://doi.org/10.1126/science.aai8830.

[8] T. Lookman, F. Alexander, K. Rajan, Information science for materials discovery and design, Springer Berlin Heidelberg, New York, NY, 2015.

[9] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[10] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2014.

[11] J.E. Gubernatis, T. Lookman, Machine learning in materials design and discovery: Examples from the present and suggestions for the future, Physical Review Materials. 2 (2018). https://doi.org/10.1103/PhysRevMaterials.2.120301.

[12] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, Npj Computational Materials. 3 (2017) 54. https://doi.org/10.1038/s41524-017-0056-5.

[13] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Physical Review Letters. 114 (2015). https://doi.org/10.1103/PhysRevLett.114.105503.

[14] J.D. Vienna, J.V. Ryan, S. Gin, Y. Inagaki, Current Understanding and Remaining Challenges in Modeling Long-Term Degradation of Borosilicate Nuclear Waste Glasses, International Journal of Applied Glass Science. 4 (2013) 283–294. https://doi.org/10.1111/ijag.12050.

[15] B. Grambow, Nuclear Waste Glasses - How Durable?, ELEMENTS. 2 (2006) 357–364. https://doi.org/10.2113/gselements.2.6.357.

[16] C.M. Jantzen, K.G. Brown, J.B. Pickett, Durable Glass for Thousands of Years, International Journal of Applied Glass Science. 1 (2010) 38–62. https://doi.org/10.1111/j.2041-1294.2010.00007.x.

[17] M. Collin, M. Fournier, P. Frugier, T. Charpentier, M. Moskura, L. Deng, M. Ren, J. Du, S. Gin, Structure of International Simple Glass and properties of passivating layer formed in circumneutral pH conditions, Npj Materials Degradation. 2 (2018) 4. https://doi.org/10.1038/s41529-017-0025-y.

[18] H.C. Helgeson, W.M. Murphy, P. Aagaard, Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions. II. Rate constants, effective surface area, and the hydrolysis of feldspar, Geochimica et Cosmochimica Acta. 48 (1984) 2405–2432. https://doi.org/10.1016/0016-7037(84)90294-1.

[19] R.H. Doremus, Diffusion-controlled reaction of water with glass, Journal of Non-Crystalline Solids. 55 (1983) 143–147. https://doi.org/10.1016/0022-3093(83)90014-5.

[20] J.K. Christie, R.I. Ainsworth, N.H. de Leeuw, Investigating structural features which control the dissolution of bioactive phosphate glasses: Beyond the network connectivity, Journal of Non-Crystalline Solids. (n.d.). https://doi.org/10.1016/j.jnoncrysol.2015.01.016.

[21] M. Bauchy, Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: A review, Computational Materials Science. 159 (2019) 95–102. https://doi.org/10.1016/j.commatsci.2018.12.004.

[22] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[23] A.K. Varshneya, Fundamentals of Inorganic Glasses, Academic Press Inc, 1993.

[24] J.P. Hamilton, C.G. Pantano, Effects of glass structure on the corrosion behavior of sodium-aluminosilicate glasses, Journal of Non-Crystalline Solids. 222 (1997) 167–174. https://doi.org/10.1016/S0022-3093(97)90110-1.

[25] B.O. Mysen, P. Richet, Silicate Glasses and Melts: Properties and Structure, Elsevier, 2005.

[26] J.P. Hamilton, Corrosion behavior of sodium aluminosilicate glasses and crystals, 1999.

[27] M.-F. Li, X.-P. Tang, W. Wu, H.-B. Liu, General models for estimating daily global solar radiation for different solar radiation zones in mainland China, Energy Conversion and Management. 70 (2013) 139–148. https://doi.org/10.1016/j.enconman.2013.03.004.

[28] J.C. Mauro, Topological constraint theory of glass, American Ceramic Society Bulletin. 90 (n.d.) 7.

[29] J.C. Phillips, Topology of covalent non-crystalline solids .1. Short-range order in chalcogenide alloys, J. Non-Cryst. Solids. 34 (1979) 153–181. https://doi.org/10.1016/0022-3093(79)90033-4.

[30] T. Oey, K.F. Frederiksen, N. Mascaraque, R. Youngman, M. Balonis, M.M. Smedskjaer, M. Bauchy, G. Sant, The role of the network-modifier's field-strength in the chemical durability of aluminoborate glasses, Journal of Non-Crystalline Solids. 505 (2019) 279–285. https://doi.org/10.1016/j.jnoncrysol.2018.11.019.

[31] T. Oey, A. Kumar, I. Pignatelli, Y. Yu, N. Neithalath, J.W. Bullard, M. Bauchy, G. Sant, Topological controls on the dissolution kinetics of glassy aluminosilicates, J Am Ceram Soc. 100 (2017) 5521–5527. https://doi.org/10.1111/jace.15122.

[32] M. Bauchy, Structural, vibrational, and elastic properties of a calcium aluminosilicate glass from molecular dynamics simulations: The role of the potential, The Journal of Chemical Physics. 141 (2014) 024507. https://doi.org/10.1063/1.4886421.

[33] Q.J. Zheng, R.E. Youngman, C.L. Hogue, J.C. Mauro, M. Potuzak, M.M. Smedskjaer, Y.Z. Yue, Structure of boroaluminosilicate glasses: Impact of [Al 2 O 3]/[SiO 2] ratio on the structural role of sodium, Physical Review B. 86 (2012). https://doi.org/10.1103/PhysRevB.86.054203.

[34] J.P. Hamilton, C.G. Pantano, S.L. Brantley, Dissolution of albite glass and crystal, Geochimica et Cosmochimica Acta. 64 (2000) 2603–2615. https://doi.org/10.1016/S0016-7037(00)00388-4.

[35] J.P. Hamilton, S.L. Brantley, C.G. Pantano, L.J. Criscenti, J.D. Kubicki, Dissolution of nepheline, jadeite and albite glasses: toward better models for aluminosilicate dissolution, Geochimica et Cosmochimica Acta. 65 (2001) 3683–3702. https://doi.org/10.1016/S0016-7037(01)00724-4.

[36] J.D. Vienna, J.J. Neeway, J.V. Ryan, S.N. Kerisit, Impacts of glass composition, pH, and temperature on glass forward dissolution rate, Npj Materials Degradation. 2 (2018) 22. https://doi.org/10.1038/s41529-018-0042-5.

[37] I. Pignatelli, A. Kumar, M. Bauchy, G. Sant, Topological Control on Silicates' Dissolution Kinetics, Langmuir. 32 (2016) 4434–4439. https://doi.org/10.1021/acs.langmuir.6b00359.

[38] E.M. Pierce, E.A. Rodriguez, L.J. Calligan, W.J. Shaw, B. Pete McGrail, An experimental study of the dissolution rates of simulated aluminoborosilicate waste glasses as a function of pH and temperature under dilute conditions, Applied Geochemistry. 23 (2008) 2559–2573. https://doi.org/10.1016/j.apgeochem.2008.05.006.

[39] N. Mascaraque, M. Bauchy, J.L.G. Fierro, S.J. Rzoska, M. Bockowski, M.M. Smedskjaer, Dissolution Kinetics of Hot Compressed Oxide Glasses, J. Phys. Chem. B. 121 (2017) 9063–9072. https://doi.org/10.1021/acs.jpcb.7b04535.

[40] P. Aagaard, H.C. Helgeson, Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions; I, Theoretical considerations, Am J Sci. 282 (1982) 237–285. https://doi.org/10.2475/ajs.282.3.237.

[41] J.C. Phillips, Topology of covalent non-crystalline solids II: Medium-range order in chalcogenide alloys and As-Si-Ge, Journal of Non-Crystalline Solids. 43 (1981) 37–77. https://doi.org/10.1016/0022-3093(81)90172-1.

[42] M.M. Smedskjaer, J.C. Mauro, Y. Yue, Prediction of Glass Hardness Using Temperature-Dependent Constraint Theory, Phys. Rev. Lett. 105 (2010) 115503. https://doi.org/10.1103/PhysRevLett.105.115503.

[43] M. Bauchy, B. Wang, M. Wang, Y. Yu, M.J. Abdolhosseini Qomi, M.M. Smedskjaer, C. Bichara, F.-J. Ulm, R. Pellenq, Fracture toughness anomalies: Viewpoint of topological constraint theory, Acta Materialia. 121 (2016) 234–239. https://doi.org/10.1016/j.actamat.2016.09.004.

[44] M. Bauchy, M. Wang, Y. Yu, B. Wang, N.M.A. Krishnan, E. Masoero, F.-J. Ulm, R. Pellenq, Topological Control on the Structural Relaxation of Atomic Networks under Stress, Phys. Rev. Lett. 119 (2017) 035502. https://doi.org/10.1103/PhysRevLett.119.035502.

[45] P.K. Gupta, J.C. Mauro, Composition dependence of glass transition temperature and fragility. I. A topological model incorporating temperature-dependent constraints, The Journal of Chemical Physics. 130 (2009) 094503-094503–8. https://doi.org/doi:10.1063/1.3077168.

[46] J.C. Mauro, P.K. Gupta, R.J. Loucks, Composition dependence of glass transition temperature and fragility. II. A topological model of alkali borate liquids, The Journal of Chemical Physics. 130 (2009) 234503-234503–8. https://doi.org/doi:10.1063/1.3152432.

[47] K. Yang, B. Yang, X. Xu, C. Hoover, M.M. Smedskjaer, M. Bauchy, Prediction of the Young's modulus of silicate glasses by topological constraint theory, Journal of Non-Crystalline Solids. 514 (2019) 15–19. https://doi.org/10.1016/j.jnoncrysol.2019.03.033.

[48] I. Pignatelli, A. Kumar, K.G. Field, B. Wang, Y. Yu, Y. Le Pape, M. Bauchy, G. Sant, Direct Experimental Evidence for Differing Reactivity Alterations of Minerals following Irradiation: The Case of Calcite and Quartz, Scientific Reports. 6 (2016) 20155. https://doi.org/10.1038/srep20155.

[49] T. Oey, Y.-H. Hsiao, E. Callagon, B. Wang, I. Pignatelli, M. Bauchy, G.N. Sant, Rate controls on silicate dissolution in cementitious environments, RILEM Technical Letters. 2 (2017) 67–73. https://doi.org/10.21809/rilemtechlett.2017.35.

[50] Y.-H. Hsiao, E.C. La Plante, N.M.A. Krishnan, Y. Le Pape, N. Neithalath, M. Bauchy, G. Sant, Effects of Irradiation on Albite's Chemical Durability, J. Phys. Chem. A. 121 (2017) 7835–7845. https://doi.org/10.1021/acs.jpca.7b05098.

[51] N. Mascaraque, M. Bauchy, M.M. Smedskjaer, Correlating the Network Topology of Oxide Glasses with their Chemical Durability, J. Phys. Chem. B. 121 (2017) 1139–1147. https://doi.org/10.1021/acs.jpcb.6b11371.

[52] Y.-H. Hsiao, E.C. La Plante, N.M.A. Krishnan, H.A. Dobbs, Y. Le Pape, N. Neithalath, M. Bauchy, J. Israelachvili, G. Sant, Role of Electrochemical Surface Potential and Irradiation on Garnet-Type Almandine's Dissolution Kinetics, J. Phys. Chem. C. 122 (2018) 17268–17277. https://doi.org/10.1021/acs.jpcc.8b04459.

[53] E. Aragones, I. Gilboa, A. Postlewaite, D. Schmeidler, Accuracy vs. Simplicity: A Complex Trade-Off, SSRN Electronic Journal. (2002). https://doi.org/10.2139/ssrn.332382.

[54] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological). 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[55] M.W. Gardner, S.R. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmospheric Environment. 32 (1998) 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

[56] C.E. Rasmussen, C.K.I. Williams, Gaussian processes for machine learning, 3. print, MIT Press, Cambridge, Mass., 2008.

Section C. Integration of Machine Learning and Simulations:

Toward Next-Generation Materials Modeling

C1. Toward More Informative Data-Driven Machine Learning:

Machine Learning Informed by Differentiable Simulations

Chapter 6. End-to-End Differentiability and Tensor Processing Unit Computing to Accelerate Materials' Inverse Design

6.1 Introduction

Numerical simulations have transformed the way we design materials [1]. For instance, density functional theory and molecular dynamics excel at predicting the properties of materials based on the knowledge of their composition and atomic structure [2,3]. This makes it possible to replace costly trial-and-error experiments by simulations so as to screen in silico promising materials [4]. However, numerical simulations are of limited help to tackle "inverse design" problems (i.e., identifying an optimal material featuring optimal properties within a given design space) [5–7]. Indeed, although numerical simulations are typically faster and cheaper than experiments, their computational burden usually prevents a thorough exploration of the design space (e.g., the systematic exploration of all possible materials' compositions) [8]. In addition, traditional numerical simulations are usually not differentiable, which prevents their seamless integration with gradient-based optimization methods [9,10]. These limitations—which are reminiscent of the state of machine learning before automatic differentiation became popular [11]—have limited the use of numerical simulations in inverse design pipelines [12].

To address this issue, it is common to replace simulations by a differentiable surrogate predictor machine learning model, which aims to approximately interpolate the mapping between design space parameters (e.g., the material's structure) and the target property of interest [7,12,13]. Following this approach, Generative Networks (GNs) [5] have been used for inverse design application using, for instance, autoencoders [14], generative adversarial networks [15], or generative inverse design networks [12]. The generator can then be combined with the differentiable surrogate predictor in the same pipeline so as to be trained by gradient

backpropagation [12,16,17]. However, this approach can result in difficulties associated with the fact that the generator and predictor must both be trained, either simultaneously or sequentially. In addition, the ability of the generator to discover new unknown, potentially non-intuitive material designs (i.e., which are very different from those in the training set) is often limited by the accuracy and generalizability of the surrogate predictor [5–7].

Then the question is, can we avoid using surrogate predictor? With the recent expansion of automatic differentiation technologies [18,19], differentiable programming platforms—such as TensorFlow [20], JAX [21], and TaiChi [22]—are rapidly developing and getting attention for differentiable simulation applications [23,24], including molecular dynamics [10,11] and robotic dynamics [25]. However, as differentiable programming remains largely unexplored in material simulations, the potential of directly training a generator based on a differentiable predictor has received less attention.

Here, to address the challenges facing surrogate predictor, we introduce a deep generative pipeline that combines an end-to-end differentiable simulator with a generator model. We illustrate the power of this approach by taking the example of the inverse design of a porous matrix featuring targeted sorption isotherm—wherein the sorption isotherm corresponds here to the amount of adsorbed liquid water in the porous structure as a function of relative humidity. This is enabled by the implementation of an end-to-end differentiable lattice-based density functional theory code in TensorFlow [20,21]. We show that the trained generative model is able to successfully generate porous structures with arbitrary sorption curves. Moreover, this generator-simulator pipeline leverages for the first time the power of tensor processing units (TPU)—an emerging family of dedicated chips [26], which, although they are specialized in deep learning, are flexible enough

for intensive scientific simulations. This approach holds promise to accelerate the inverse design of materials with tailored properties and functionalities.

6.2 Methods

6.2.1 Lattice density function theory (LDFT) of sorption

We consider a simplified model of porous matrix by using a square *N*-by-*N* lattice (see Fig. 6-1a). In this lattice, the state of each pixel *i* is given by the knowledge of (η_i, ρ_i), where $\eta_i = 0$ and 1 indicate that the pixel is filled with solid or is a pore, respectively, and ρ_i is the density of water in the pore upon increasing relative humidity (RH). $\rho_i = 0$ and 1 denote that the pore is fully empty or saturated with water, respectively. According to LDFT [27,28], the equilibrium $\{\rho_i\}$ at a given RH is computed by minimizing the configuration's grand potential function $\Omega(\{\rho_i\})$:

$$\Omega(\{\rho_i\}) = kT\sum_i[\rho_i \ln(\rho_i) + (\eta_i - \rho_i)\ln(\eta_i - \rho_i)] - w_{\mathrm{ff}}\sum_{\langle ij\rangle}\rho_i\rho_j - w_{\mathrm{mf}}\sum_{\langle ij\rangle}\big[\rho_i(1-\eta_j) + \rho_j(1-\eta_i)\big] - \mu\sum_i \rho_i \qquad \textbf{Eq. (6-1)}$$

where *k* is the Boltzmann constant, *T* is temperature (here, *T* = 298 K), w_{ff} is interaction energy between two neighboring pixels that are filled with water, w_{mf} is the interaction energy between a pixel filled with water and a neighboring pixel filled with solid, $\langle ij\rangle$ indicates the sums are restricted to distinct nearest neighbor pairs (note that, to avoid any surface effect, periodic boundary conditions are applied), μ is the chemical potential (which depends on RH). μ is calculated by $\mu = \mu_{sat} + kT\ln(\mathrm{RH})$, where μ_{sat} is the chemical potential of water at saturated state and can be estimated as $\mu_{sat} = w_{\mathrm{ff}} \times c/2$ (here, the coordination number c = 4). w_{ff} is derived from the critical temperature of water (T_c = 647 K) and $w_{\mathrm{ff}} = 4kT_c/c$. w_{mf} is calculated from the interaction ratio parameter $y = w_{\mathrm{mf}}/w_{\mathrm{ff}}$, where $y > 1$ indicates a hydrophilic substrate

and $y < 1$ a hydrophobic substrate (here, $y = 1.5$). By minimizing Eq. (6-1) with respect to $\{\rho_i\}$, the solution of equilibrium $\{\rho_i\}$ is rewritten as an iteration loop of Eq. (6-2) until convergence:

$$\rho_i = \frac{\eta_i}{1+e^{-\{\mu+\sum_{j/i}[w_{\mathrm{ff}}\rho_j+w_{\mathrm{mf}}(1-\eta_j)]\}/(kT)}} \quad \textbf{Eq. (6-2)}$$

where j/i are the pixel IDs of the 4 neighbors of pixel i [28], that is, the water density at a given pixel i depends on the state of its 4 neighbors—which is essentially a convolution operation—and the convergence condition is set as $1/N^2 \sum_i(\rho_i^{(t+1)} - \rho_i^{(t)}) < 10^{-10}$ between two consecutive loop t and t+1. ρ_i is calculated at each RH for RH = 0-to-100% with an increment dRH (here, dRH = 2.5%). Initially, the equilibrium $\rho_i = 0$ when RH = 0. At each increment K, the equilibrium water density values $\{\rho_i\}^{K\text{th}}$ at RH = K × dRH serve as starting configuration to calculate $\{\rho_i\}^{K+1}$ at the subsequent step K+1 by iteratively applying Eq. (6-2) until a convergence in the $\{\rho_i\}$ values is obtained. Finally, the water sorption isotherm $\{\rho_w\}_{1\times 40}$ is obtained by calculating the average pore water density $\rho_w = \langle\rho_i\rangle = 1/N^2 \sum_i(\rho_i)$ at each of the 40 RH increments. More detailed descriptions of LDFT of sorption can be found in Ref. [28].

6.2.2 End-to-end differentiable implementation of LDFT

For a sequence of RH = 0-to-100% with an increment dRH = 2.5%, the numerical simulation of sorption at each of the 40 RH increments can be briefly described as an iterative loop of Eq. (6-2) until convergence. Then the key aspect to implement a differentiable simulation lies in decomposing Eq. (6-2) into a series of mathematical operations that can be implemented as differentiable computation layers in TensorFlow (see Fig. 6-1b). Here, we decompose Eq. (6-2) into three layers, namely, (i) the input layer, (ii) the convolution layer, and (iii) the output layer. The input layer consists of three parallel layers associated with three input matrices, respectively, where one input matrix $\{\eta_i\}_{N\times N}$ is fed into the output layer, and the other two matrices $\{\rho_i\}_{N\times N}$

and $\{1-\eta_i\}_{N\times N}$ are fed into two parallel convolution layers. Then the two convolution layers conduct the convolution operation $\sum_{j/i}[w_{\mathrm{ff}}\rho_j]$ and $\sum_{j/i}[w_{\mathrm{mf}}(1-\eta_j)]$, respectively, wherein *j*/*i* indicates 4 neighbors of pixel *i* (note that, to avoid any surface effect, periodic boundary conditions are applied [28]). Finally, the two convolution outcomes (denoted as C1 and C2) together with the input matrix $\{\eta_i\}_{N\times N}$ is fed into the output layer that conducts the remaining mathematical operation of Eq. (6-2), namely, $\rho_i = \frac{\eta_i}{1+e^{-\{\mu+C1+C2\}/(kT)}}$. Importantly, since all layers share the feature of automatic differentiation in TensorFlow, the gradient of each layer can back propagate to enable end-to-end differentiation. At a given RH, we repeat the block (i.e., the decomposed layers of Eq. (6-2)) for *M* times in series (here, *M* = 100), which is equivalent to iteratively solving Eq. (6-2) until a convergence in the water density is achieved.

6.2.3 Structure of the generator-simulator pipeline

The generator is designed as a dual, parallel deconvolution-block structures (see Fig. 6-2a), where each block is fed with half of the input sorption isotherm $\{\rho_w\}_{1\times 40}$. In detail, the low- and high-RH block is fed with the first and second half of the input $\{\rho_w\}_{1\times 40}$, i.e., a 1-by-20 array each. Then the two blocks show the same structure, which consists of 4 layers in series, that is, (i) a fully connected dense neural layer (DENSE) that contains $20 \times 20 \times 64 = 25600$ neurons and outputs a 1-by-25600 array; (ii) a reshape layer (RESHAPE) that transforms the one dimensional 1-by-25600 array into a three dimensional 20-by-20-by-64 array; (iii) a deconvolution layer (DECONV) that contains 64 channels with a 20×20 filter size and outputs a three dimensional 20-by-20-by-64 array; (iv) a convolution layer (CONV) that contains 1 channel with a 3×3 filter size and outputs a two-dimensional 20-by-20 array. The activation function of each layer is set as "ReLU" function, and batch normalization has been applied to the output of each layer to

accelerate the training process [29]. Finally, the two 20-by-20 array obtained from the low- and high-RH block—which are denoted here as low- and high-RH activation, respectively—are concatenated in parallel and are fed into the generator's output layer, namely, a convolution layer that contains 1 channel, uses a 3×3 filter size and a "binary sigmoid" activation function, and outputs a 20-by-20 prediction gird $\{\eta_i\}_{20\times20}$. The generator output $\{\eta_i\}_{20\times20}$ is then fed into the differentiable simulator for validation. This configuration would go through $M = 100$ consecutive blocks of TensorFlow-based layers (i.e., the decomposed operations of Eq. (6-2) programmed in TensorFlow, see last section) at each of the 40 RH increments to obtain the output sorption isotherm $\{\rho_w\}_{1\times40}$.

6.2.4 Preparation of training and test sets

The training set contains 6,400,000 target sorption curves. These curves are generated automatically from a self-defined generative function. This function aims to produce as many as possible curves that are monotonically non-decreasing but vary differently in terms of trend, convexity, and value. Although this generative curve are not real sorption isotherms, they possess most important features of real sorption curves and cover all possible variations of real sorption isotherms. There are different ways to define the generative function. Here we propose one type of generative function that satisfy the above requirements. This function generates 20% "stepwise" curves and 80% "anchor-based" curves. By discretizing the curve as a one dimensional 1-by-40 array $\{\rho_w\}_{1\times40}$, the "stepwise" curves are designed as an array where the first n elements = 0 and the last $(40 - n)$ elements = 1, where the integer n is uniformly randomized from 1 to 39. The "anchor-based" curves are designed by first defining an "anchor" element from the 1-by-40 array, where the anchor is the n-th element, and we uniformly randomize its index n from 1 to 39 and its

value A from 0 to 1. Then, regarding all the elements before the anchor, the increment D of their value between two consecutive elements can be expressed as $D = e^R / \sum_n e^R \times A$, where R is a random number sampled from a normal distribution with a zero mean and a standard deviation of σ (here, σ = 4). Similarly, regarding all the elements after the anchor, the increment D of their value between two consecutive elements can be expressed as $D = e^R / \sum_{40-n} e^R \times (1 - A)$. Both the "stepwise" and "anchor-based" curves can be generated efficiently to create a large training set covering a diverse population of sorption curves. Finally, the test set are real sorption curves to evaluate the generator's prediction accuracy. Here, we create a test set that contains 8769 real curves. These curves are generated by the sorption simulator using a large set of grids (see Fig. 6-1c-ii), which includes 8769 diverse and random grid patterns.

6.2.5 Training of the generator-simulator pipeline

In the training process, we first set the grid size N = 20 and the batch size = 64. Then we train the pipeline for 100 epochs and each epoch contains 1000 batches. The loss function used herein is the percentage loss L between the forward output and the reference target curve (see Fig. 6-2b), that is, the area between the forward curve and the reference curve. It is worthwhile to point out that, since both the solid phase and pore phase in a porous matrix shows some continuity within their phase, some regularization term can be applied to the training process to simultaneously accelerate the training and improve the prediction accuracy [29]. Here, the regularization term designed for the generator output is defined as $\sum_i^{N^2} \sum_{j/i} |\eta_i - \eta_j| / 4N^2$, which panelizes a solid site neighbored by a pore, or vice versa. In other words, the generator's output gird would favor continuous solid phase or continuous pore phase. Then we select the Stochastic Gradient Descent (SGD) optimizer to minimize the loss function [29]. The momentum is set as 0.9 to accelerate

gradient descent [20]. The learning rate is initially set as 10^{-2} and gradually decays by a factor of 0.1 after a patience of 10 epochs [20]. Finally, a validation step is applied to the pipeline after each training epoch using the test set of 8769 sorption curves (see Fig. 6-2c).

6.3 Results

6.3.1 End-to-end differentiable simulator

We first focus on the end-to-end differentiable implementation of water sorption simulation for a target porous matrix. We consider as a toy model a square N-by-N lattice, wherein each pixel i of the grid can be filled with solid or be a pore (see Figure 6-1a). Initially empty pixel can then be filled with water upon increasing relative humidity (RH). In a given configuration, the state of each pixel i is given by the knowledge of (η_i, ρ_i), where $\eta_i = 0$ and 1 indicate that the pixel is filled with solid or is a pore, respectively, and ρ_i is the density of water in the pore ($\rho_i = 0$ and 1 denote that the pore is fully empty or saturated with water, respectively). The equilibrium fraction of water in each pore at given temperature T and RH is then solved by lattice density functional theory (LDFT) [27,28]. Based on this formalism, the water density ρ_i at a given pixel i is given by Eq. (6-2), which is essentially a convolution operation (see Methods section). At fixed RH, the equilibrium fraction of water is then determined by iteratively applying Eq. (6-2) on each pixel until a convergence in the $\{\rho_i\}$ values is obtained. The sorption of water in the porous matrix is then iteratively simulated by computing the equilibrium values of $\{\rho_i\}$ for RH = 0-to-100% with an increment dRH. At each increment K, the equilibrium water density values $\{\rho_i\}^{K\text{th}}$ at RH = $K \times$ dRH serve as starting configuration to calculate $\{\rho_i\}^{K+1}$ at the subsequent step K+1. More details of the numerical simulations can be found in the Methods section.

Such (LDFT) simulations are traditionally not differentiable. Here, to address this limitation, we decompose Eq. (6-2) into a series of mathematical operations that can be implemented as differentiable computation layers in TensorFlow (Figure 6-1b). For instance, the CONV layer represents the convolution operation in Eq. (6-2)—i.e., one of the operations that can be efficiently performed by TPUs. This block is then repeated into M convolutional layers, which is equivalent to iteratively solving Eq. (6-2) until a convergence in the water density is achieved. More details of the differentiable simulations can be found in the Methods section.

We now evaluate the accuracy of differentiable simulator with respect to the conventional (ground-truth but undifferentiable) simulator. Figure 6-1c-i shows a comparison between sorption curve computed by the ground-truth simulator and by the differentiable counterpart for the porous matrix shown in Figure 6-1a, where the accuracy of differentiable simulator is characterized by the percentage loss L of sorption curve obtained from differentiable simulator, that is, the area between this curve and the ground-truth reference curve. We then evaluate the average percentage loss of the differentiable simulator by using a large validation set of porous matrices. The validation set contains 8769 grids associated with a diverse population of reference sorption curves (see Figure 6-1c-ii), as characterized by a wide distribution of the sinuosity index of reference curve S_r, where S_r is calculated as the ratio of the curvilinear length along the curve over the straight-line length between end points of the curve. Using this validation set, Figure 6-1c-iii shows the accuracy of the TensorFlow-based simulator as a function of the number of convolution layers M. We find that a larger M value helps the convergence of water density and thus improves the simulation accuracy, and in our case, $M = 100$ offers satisfactory accuracy. Finally, Figure 6-1c-iv provides the average percentage loss $\langle L \rangle$ as a function of S_r at $M = 100$. We find that $\langle L \rangle$ for all ranges of S_r is around 0.36%, illustrating the differentiable simulators is as accurate as the ground-

truth simulator. Overall, by reformulating the LDFT simulation into a succession of convolutional layers, this approach enables end-to-end differentiability and TPU acceleration.

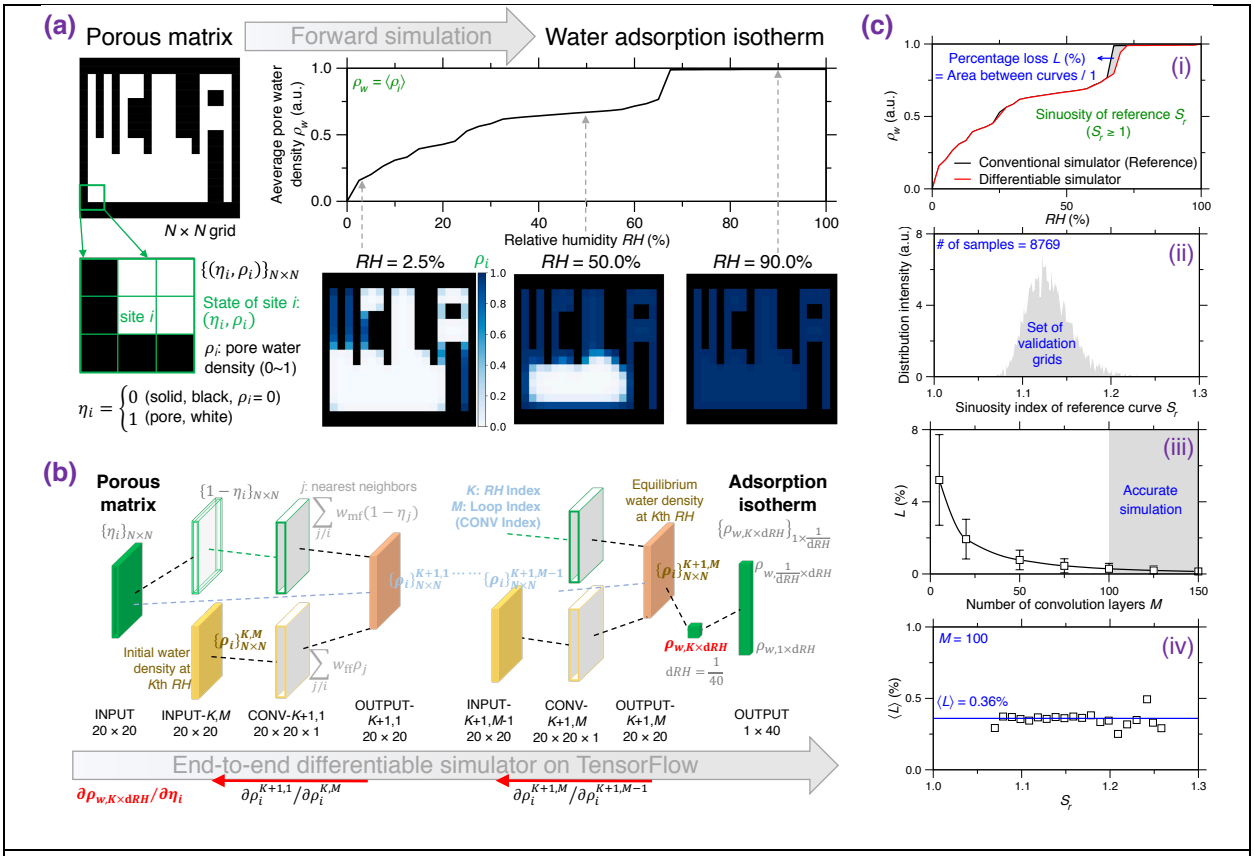


Figure 6-1: End-to-end differentiable simulation of water adsorption in porous materials. **(a)** Illustration of the numerical water sorption simulation for a target porous matrix. The porous matrix is represented by a *N*-by-*N* grid, wherein each pixel *i* of the grid can be filled with solid ($\eta_i = 0$) or be a pore ($\eta_i = 1$). ρ_i is the density of water in the pore. ρ_i = 0 and 1 denote that the pore is fully empty or saturated with water, respectively. ρ_i is calculated at each relative humidity (RH) for RH = 0-to-100% with an increment dRH. At each increment *K*, the equilibrium water density values $\{\rho_i\}^{K\text{th}}$ at RH = K × dRH serve as starting configuration to calculate $\{\rho_i\}^{K+1}$ at the subsequent step K+1, where the equilibrium fraction of water is determined by iteratively applying Eq. (6-2) on each pixel until a convergence in the $\{\rho_i\}$

values is obtained. **(b)** End-to-end differentiable reformation of the sorption simulation as a series of differentiable computation layers in TensorFlow. Each layer is a mathematical operation by decomposing Eq. (6-2), where CONV layer represents the convolution operation in Eq. (6-2). This block is then repeated into M convolutional layers, which is equivalent to iteratively solving Eq. (6-2) until a convergence in the water density is achieved. **(c-i)** Comparison between the sorption curve ground-truth (undifferentiable) sorption simulator and its reformulated differentiable counterpart for the porous matrix shown in panel (a), which defines the percentage loss L. **(c-ii)** Distribution of the sinuosity index of reference curve (i.e., ground-truth sorption curve) S_r for 8769 validation grids. S_r is calculated as the ratio of the curvilinear length along the curve over the straight-line length between end points of the curve. **(c-iii)** Average percentage loss as a function of the number of convolution layers M. The grey window ($M \geq 100$) indicates the range where the differentiable simulator is as accurate as the ground-truth simulator. **(c-iv)** Average percentage loss $\langle L \rangle$ as a function of S_r at $M = 100$. The blue line represents the average percentage loss for 8769 validation grids.

6.3.2 Architecture of the generator-simulator pipeline

Now we are in the stage to train a generative model directly from the end-to-end differentiable simulator presented above. We first present the architecture of our generative model and its integration with the differentiable simulator. Figure 6-2a shows the architecture of the generator-simulator pipeline. In detail, the training pipeline takes as inputs the sorption isotherm curves of the training set, which are transformed into porous matrices by the generator. The generated grids are then fed to the differentiable simulator to compute the "real" sorption curve of the generated porous matrices. In detail, the generator is designed as a dual, parallel deconvolution-

block structure, where each block is fed with half of the input curve to decrease the generator complexity. These two blocks aim to specifically generate small and large pores, which are saturated with water at low and large RH, respectively. More details about the architecture of the generator-simulator pipeline are provided in the Methods section. Since each layer of the pipeline is differentiable, the generator can then be optimized by gradient backpropagation in TensorFlow so as to minimize the difference between the input and output sorption curves. Note that, here, the convolutional layers of the simulator are hard-coded with fixed weights and, hence, are not optimized. This is key advantage of our approach since it avoids difficulties arising from the simultaneous optimization of the generator and predictor in traditional implementations of generative pipelines.

6.3.3 Training acceleration by Tensor Processing Unit computing

In this section, we focus on the training of the generator-simulator pipeline to minimize the difference between input and output sorption curve, i.e., the loss function L (see Fig. 6-2b). During the training process, a grid size of 20×20 yields about 7 million parameters to be optimized for the generator, while the simulator comprises about 4000 convolution layers to compute. Here, the generator is trained based on a training set of 6,400,000 sorption isotherm curves and then subsequently evaluated based on a test set of 8,769 curves. More details of the training and test sets can be found in the Methods section. Figure 6-2c shows the evolution of the test set loss function L as a function of the number of training epochs, wherein the batch size is set as 64 and each epoch contain 1000 batches. We find that the accuracy of the generator plateaus after 50 epochs (which corresponds to a training size of 3,200,000). Figure 6-2d further shows the average loss function as a function of the sinuosity index of reference curve (i.e., input sorption curve) S_r

in test set at epoch = 100. We find that the generator exhibits an average prediction loss of 3%, which is here considered very good (see below).

Considering the large depth of the simulator and the number of parameters to be optimized in the generator, the training process comes with a significant computational cost. To mitigate this issue, as a pioneering experiment, the training is conducted on TPUs [26]. TPU is a family of dedicated chips that assemble different computing units for machine learning applications [30]. Figure 6-2e-i shows a schematic of the TPU computing system composed of both software and hardware architecture, where TensorFlow is a software used to compile program ready for TPU computing on TPU chip. In contrast to general purposes processors (i.e., CPUs and GPUs), TPUs are specifically designed as matrix processors thanks to their matrix unit (MXU) [31,32]. Although TPUs have been extensively used for deep learning, their application to numerical simulations has thus far remained limited an Ising model [33]. However, TPUs exhibit enough flexibility to have the potential to accelerate a broader range of computations. Figure 6-2e-ii show the training time per batch as a function of both grid size and batch size on a TPU-v2 chip with 8 cores and 64 GB memory [26]. The computational performance is compared with the training time yielded by a NVIDIA TITAN X GPU. All benchmarks are conducted on Google Colab using the same TensorFlow code and single precision (float32). Figure 6-2e-iii further describes the TPU and GPU training time as a function of batch size for grid size $N = 20$ and 80. We find that, especially for large grid size and batch size, the deliciated TPU hardware results in a training time that is several times faster than that offered by the GPU hardware considered herein (more than 6× faster, see Fig. 6-2e-iv). These results highlight the exciting, largely untapped potential of TPU computing in accelerating computationally-intensive scientific simulations (i.e., besides traditional deep learning applications). More details of the pipeline training can be found the Methods section.

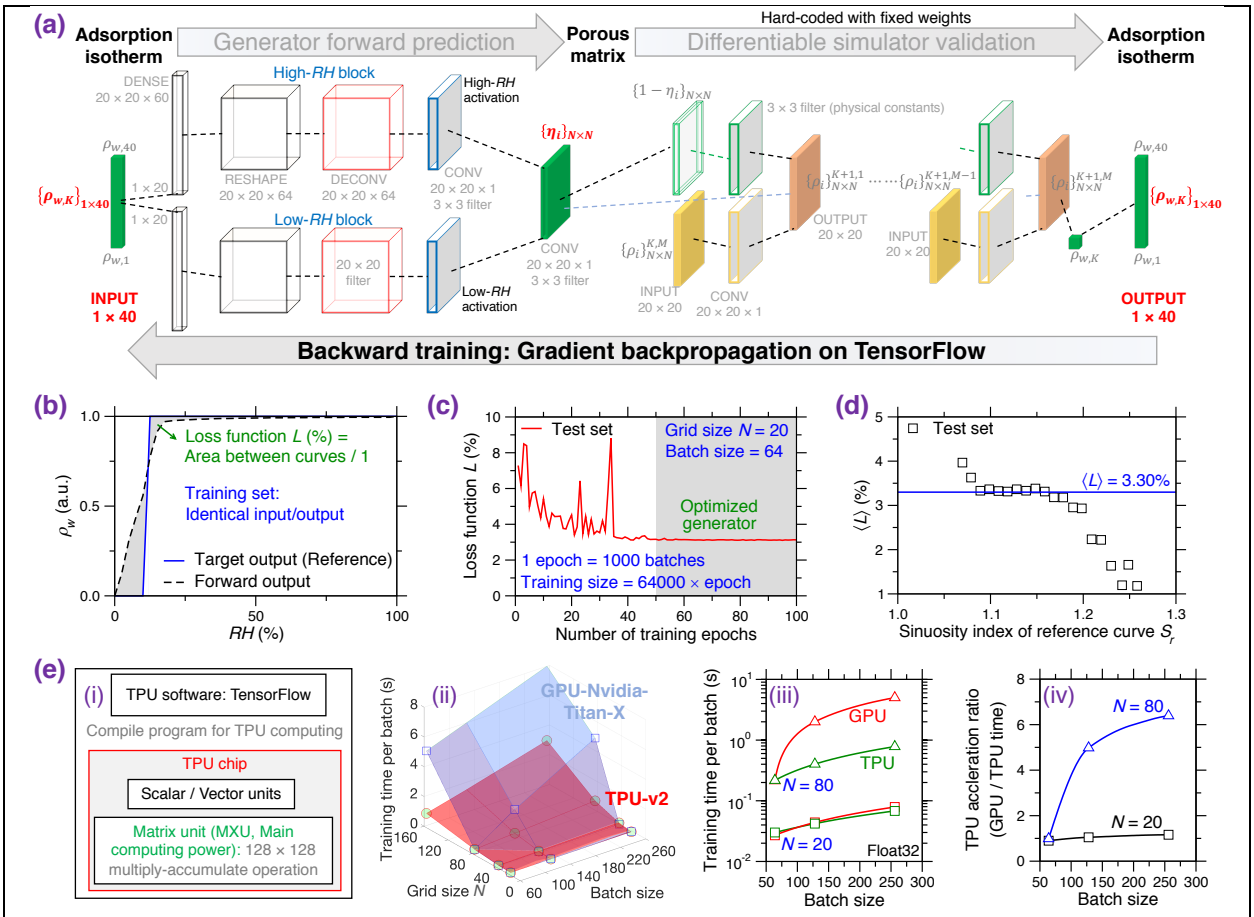


Figure 6-2: Training of the generative model by differentiable simulation and tensor processing unit (TPU) computing. **(a)** General architecture of the generator-simulator training pipeline. The generator is designed as a dual, parallel deconvolution-block structure, where each block is fed with half of the input curve $\{\rho_{w,K}\}$ that represents low- and high-RH range signal, respectively. The associated porous matrix $\{\eta_i\}$ predicted by the generator is subsequently fed to the differentiable simulator for validation. The forward output of the simulator is then compared with the targeted output—which is the same as generator input $\{\rho_{w,K}\}$—to calculate the loss function used for backward training on TensorFlow. **(b)** Loss function L (grey area) for a target output (blue line). **(c)** Evolution of the test set loss function as a function of the number of training epochs. The test set contains 8769 validation curves

(see Fig. 6-1c-ii). The plateau in the grey window indicates the generator reaches optimal prediction performance. **(d)** Average test set loss function as a function of the sinuosity index of reference curve (i.e., target output) S_r at epoch = 100. **(e-i)** Schematic of the TPU computing system composed of both software and hardware architecture, where TensorFlow is a software used to compile program ready for TPU computing on TPU chip. TPU chip is an assembly of different computing units specific for machine learning, where the main computing power arises from the matrix unit (MXU) capable of 128 × 128 multiply-accumulate operation. **(e-ii)** Comparison of the training time per batch as a function of the grid size and batch size offered by Google's TPU-v2 and an NVIDIA TITAN X GPU. All benchmarks are conducted on Google Colab using the same TensorFlow code and single precision (float32). **(e-iii)** Detailed comparison of the training time per batch between TPU and GPU as a function of batch size for grid size N = 20 and 80. **(e-iv)** TPU acceleration ratio (defined as GPU time / TPU time) as a function of batch size for grid size N = 20 and 80.

6.3.4 Accuracy of the generator

Finally, we evaluate the accuracy of the trained generator on the test set (which comprises more than 8000 target sorption isotherms). After training, we find that the generator exhibits an average prediction loss of 3% (see Figure 6-2d), which is here considered very good. Figure 6-3a offers an illustration of three porous matrices that are generated so as to present three archetypical sorption isotherms wherein: (i) full water saturation occurs at very low RH (which arises in the presence of very small pores), (ii) water saturation is delayed and occurs at very large RH (which is a consequence of large pores), and (iii) an intermediate case (with medium-size pores). Overall, we find that the generator model is able to predict realistic porous matrices, with expected length

scales for the pores. Importantly, the simulated sorption curves of the generated porous structures exhibit all the features (in terms of trend, convexity, and value) as the target sorption curves.

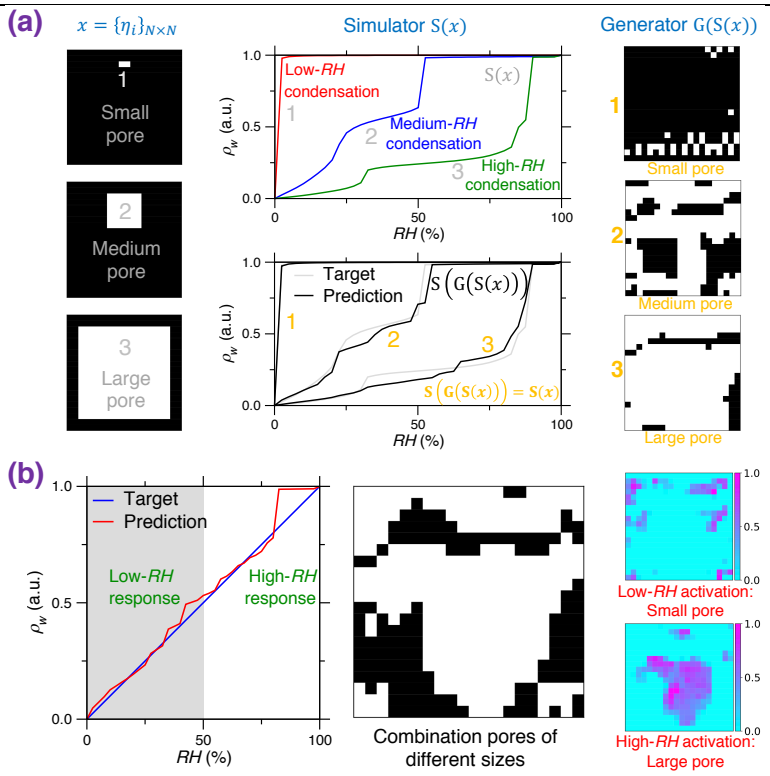


Figure 6-3: Accuracy of the generative model. **(a)** Illustration of three porous matrices that are generated so as to present three archetypical sorption isotherms associated with small, medium, and large pores. **(b)** Porous matrix generated for a target sorption curve $y = x$. The activation pattern of low- and high-RH block is also provided.

As a final test of the generator, we assess the ability of the generator to predict a porous structure featuring a target identity sorption curve $y = x$. This is an especially challenging test set case since (i) the sorption curve is not included in the training set, (ii) such a smooth sorption curve (with no sudden jump in water sorption) requires a complex, continuous pore size distribution, and (iii) this case corresponds to maximum degeneracy—unlike the cases of a 1-pixel or $(N-1)\times(N-$

1) pores, which present a limited number of possible solutions. Once again, we find that the generator yields a very realistic generated porous matrix, which, as expected, exhibits a combination of small, medium, and large pores (see Fig. 6-3b). Notably, the real sorption curve (computed by the simulator) of the generated porous matrix indeed exhibit a very close match with the $y = x$ target. This confirms that the generative model has learned the basic physical rules governing water sorption in porous media (e.g., small and large pores get saturated as low and high RH, etc.) and can successfully predict new unknown porous structures featuring tailored arbitrary sorption curves. In that regard, the fact that the generator is directly trained based on the simulator (rather than on surrogate model that approximates reality by learning from finite training set examples) is key to ensure that the generator is not limited by the accuracy of the predictor, or its ability to extrapolate predictions to grids it has never been exposed to during its training.

6.4 Discussion

By designing as a dual, parallel-block structure (see Fig. 6-2a), the generator shows not only high prediction accuracy but also enough simplicity and interpretability (as compared to a single, giant-block structure). After training, we find these two blocks can specifically generate small and large pores that are saturated with water at low and large RH, respectively (see Fig. 6-3b), in agreement with the basic physics of fluid sorption that small and large pores exhibit early and delayed condensation behavior, respectively [27,28]. These results *a posterior* demonstrate that the physics-informed machine learning framework would simultaneously ensure both the model simplicity and the prediction accuracy [34].

This research has several scientific and societal implications. First, this work illustrates the benefits of integrating differentiable simulations in machine learning pipelines—which is key to

accelerate the discovery of new materials. Second, our results establish TPU computing as a promising route to accelerate scientific simulations, which are ubiquitous in various applications (drug discovery by molecular dynamics, architectural design by finite element method, weather forecast predictions, etc.) [1–3]. Finally, the ability to design new porous structures with tailored sorption isotherms could leapfrog several important applications, including for CO_2 capture [35,36] and gas separation [37,38]. In addition, designing new porous structures featuring a smooth, continuous sorption isotherm (i.e., as close as possible to the $y = x$ target used herein) is important for drug delivery applications, to ensure that drugs are continuously released at a constant rate in a given environment [39,40].

6.5 Conclusions

Overall, this work establishes a robust pipeline to enable the inverse design of materials by leveraging an end-to-end differentiable simulation as predictor. The fact that the generator is directly trained based on a simulator rather than on a surrogate machine learning model is key to ensure that the generator is not limited by the accuracy or extrapolation ability of the predictor. As a key enabler of this approach, we adopt TPUs to accelerate the training of the generator by gradient backpropagation in TensorFlow. This illustrates the exciting possibilities of TPU computing to accelerate scientific numerical simulations.

6.6 References

[1] E.V. Levchenko, Y.J. Dappe, G. Ori, eds., Theory and Simulation in Physics for Materials Applications: Cutting-Edge Techniques in Theoretical and Computational Materials Science, Springer International Publishing, Cham, 2020. https://doi.org/10.1007/978-3-030-37790-8.

[2] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Materials. 4 (2016) 053208. https://doi.org/10.1063/1.4946894.

[3] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[4] E.O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery, Annual Review of Materials Research. 45 (2015) 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.

[5] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science. 361 (2018) 360–365. https://doi.org/10.1126/science.aat2663.

[6] T.W. Liao, G. Li, Metaheuristic-based inverse design of materials – A survey, Journal of Materiomics. 6 (2020) 414–430. https://doi.org/10.1016/j.jmat.2020.02.011.

[7] J. Noh, G.H. Gu, S. Kim, Y. Jung, Machine-enabled inverse design of inorganic solid materials: promises and challenges, Chem. Sci. 11 (2020) 4871–4881. https://doi.org/10.1039/D0SC00594K.

[8] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: A review, Journal of Non-Crystalline Solids: X. 4 (2019) 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[9] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[10] W. Wang, S. Axelrod, R. Gómez-Bombarelli, Differentiable Molecular Simulations for Control and Learning, ArXiv:2003.00868 [Physics, Stat]. (2020). http://arxiv.org/abs/2003.00868 (accessed April 19, 2020).

[11] S.S. Schoenholz, E.D. Cubuk, JAX, M.D.: End-to-End Differentiable, Hardware Accelerated, Molecular Dynamics in Pure Python, ArXiv:1912.04232 [Cond-Mat, Physics:Physics, Stat]. (2019). http://arxiv.org/abs/1912.04232 (accessed March 25, 2020).

[12] C.-T. Chen, G.X. Gu, Generative Deep Neural Networks for Inverse Materials Design Using Backpropagation and Active Learning, Advanced Science. 7 (2020) 1902607. https://doi.org/10.1002/advs.201902607.

[13] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, J. Hu, Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials, Npj Comput Mater. 6 (2020) 1–7. https://doi.org/10.1038/s41524-020-00352-0.

[14] D.P. Kingma, M. Welling, An Introduction to Variational Autoencoders, FNT in Machine Learning. 12 (2019) 307–392. https://doi.org/10.1561/2200000056.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014: pp. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (accessed October 3, 2020).

[16] B. Kim, S. Lee, J. Kim, Inverse design of porous materials using artificial neural networks, Science Advances. 6 (2020) eaax9324. https://doi.org/10.1126/sciadv.aax9324.

[17] Z. Liu, D. Zhu, S.P. Rodrigues, K.-T. Lee, W. Cai, Generative Model for the Inverse Design of Metasurfaces, Nano Lett. 18 (2018) 6570–6576. https://doi.org/10.1021/acs.nanolett.8b03171.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: 2017.

[19] A. Griewank, A. Walther, Evaluating Derivatives, Society for Industrial and Applied Mathematics, 2008. https://doi.org/10.1137/1.9780898717761.

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, (n.d.) 19.

[21] R. Frostig, M.J. Johnson, C. Leary, Compiling machine learning programs via high-level tracing, (n.d.) 3.

[22] Y. Hu, L. Anderson, T.-M. Li, Q. Sun, N. Carr, J. Ragan-Kelley, F. Durand, DiffTaichi: Differentiable Programming for Physical Simulation, ArXiv:1910.00935 [Physics, Stat]. (2020). http://arxiv.org/abs/1910.00935 (accessed August 20, 2020).

[23] A. Hernández, J.M. Amigó, Differentiable programming and its applications to dynamical systems, ArXiv:1912.08168 [Cs, Math]. (2020). http://arxiv.org/abs/1912.08168 (accessed August 20, 2020).

[24] F. de Avila Belbute-Peres, K. Smith, K. Allen, J. Tenenbaum, J.Z. Kolter, End-to-End Differentiable Physics for Learning and Control, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018: pp. 7178–7189. http://papers.nips.cc/paper/7948-end-to-end-differentiable-physics-for-learning-and-control.pdf (accessed June 3, 2020).

[25] Y. Hu, J. Liu, A. Spielberg, J.B. Tenenbaum, W.T. Freeman, J. Wu, D. Rus, W. Matusik, ChainQueen: A Real-Time Differentiable Physical Simulator for Soft Robotics, in: 2019 International Conference on Robotics and Automation (ICRA), 2019: pp. 6265–6271. https://doi.org/10.1109/ICRA.2019.8794333.

[26] Cloud TPU | Google Cloud, (n.d.). https://cloud.google.com/tpu (accessed October 3, 2020).

[27] E. Kierlik, P.A. Monson, M.L. Rosinberg, G. Tarjus, Adsorption hysteresis and capillary condensation in disordered porous solids: a density functional study, J. Phys.: Condens. Matter. 14 (2002) 9295–9315. https://doi.org/10.1088/0953-8984/14/40/319.

[28] E. Kierlik, M. L. Rosinberg, G. Tarjus, P. Viot, Equilibrium and out-of-equilibrium (hysteretic) behavior of fluids in disordered porous materials: Theoretical predictions, Physical Chemistry Chemical Physics. 3 (2001) 1201–1206. https://doi.org/10.1039/B008636N.

[29] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[30] Y.E. Wang, G.-Y. Wei, D. Brooks, Benchmarking TPU, GPU, and CPU Platforms for Deep Learning, (2019). https://arxiv.org/abs/1907.10701v4 (accessed September 4, 2020).

[31] T. Lu, Y.-F. Chen, B. Hechtman, T. Wang, J. Anderson, Large-Scale Discrete Fourier Transform on TPUs, ArXiv:2002.03260 [Cs]. (2020). http://arxiv.org/abs/2002.03260 (accessed September 2, 2020).

[32] F. Huot, Y.-F. Chen, R. Clapp, C. Boneti, J. Anderson, High-resolution imaging on TPUs, ArXiv:1912.08063 [Physics]. (2019). http://arxiv.org/abs/1912.08063 (accessed September 2, 2020).

[33] K. Yang, Y.-F. Chen, G. Roumpos, C. Colby, J. Anderson, High performance Monte Carlo simulation of ising model on TPU clusters, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, Denver Colorado, 2019: pp. 1–15. https://doi.org/10.1145/3295500.3356149.

[34] H. Liu, T. Zhang, N.M.A. Krishnan, M.M. Smedskjaer, J.V. Ryan, S. Gin, M. Bauchy, Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning, Npj Mater Degrad. 3 (2019) 1–12. https://doi.org/10.1038/s41529-019-0094-1.

[35] G. Sneddon, A. Greenaway, H.H.P. Yiu, The Potential Applications of Nanoporous Materials for the Adsorption, Separation, and Catalytic Conversion of Carbon Dioxide, Advanced Energy Materials. 4 (2014) 1301873. https://doi.org/10.1002/aenm.201301873.

[36] K. Sumida, D.L. Rogow, J.A. Mason, T.M. McDonald, E.D. Bloch, Z.R. Herm, T.-H. Bae, J.R. Long, Carbon Dioxide Capture in Metal–Organic Frameworks, Chemical Reviews. 112 (2012) 724–781. https://doi.org/10.1021/cr2003272.

[37] P.G. Boyd, Y. Lee, B. Smit, Computational development of the nanoporous materials genome, Nature Reviews Materials. 2 (2017) 17037. https://doi.org/10.1038/natrevmats.2017.37.

[38] P. Nugent, Y. Belmabkhout, S.D. Burd, A.J. Cairns, R. Luebke, K. Forrest, T. Pham, S. Ma, B. Space, L. Wojtas, M. Eddaoudi, M.J. Zaworotko, Porous materials with optimal adsorption thermodynamics and kinetics for CO2 separation, Nature. 495 (2013) 80–84. https://doi.org/10.1038/nature11893.

[39] E.J. Anglin, L. Cheng, W.R. Freeman, M.J. Sailor, Porous silicon in drug delivery devices and materials, Advanced Drug Delivery Reviews. 60 (2008) 1266–1277. https://doi.org/10.1016/j.addr.2008.03.017.

[40] P. Horcajada, T. Chalati, C. Serre, B. Gillet, C. Sebrie, T. Baati, J.F. Eubank, D. Heurtaux, P. Clayette, C. Kreuz, J.-S. Chang, Y.K. Hwang, V. Marsaud, P.-N. Bories, L. Cynober, S. Gil, G. Férey, P. Couvreur, R. Gref, Porous metal–organic-framework nanoscale carriers as a potential platform for drug delivery and imaging, Nature Mater. 9 (2010) 172–178. https://doi.org/10.1038/nmat2608.

Section C. Integration of Machine Learning and Simulations: Toward Next-Generation Materials Modeling

C2. Toward Less Complex Physics-Dirven Simulation: Machine Learning-Aided Development of Empirical Forcefields

Chapter 7. Parameterization of Empirical Forcefields for Glassy Silica Using Machine Learning

7.1 Introduction

Classical molecular dynamics (MD) simulation is an effective tool to access the atomic structure of glass, which usually remains invisible from traditional experimental techniques [1–3]. In turn, better understanding the atomic structure of glasses is key to decipher their genome, that is, to understand how their composition and structure control their engineering properties [4]. However, the accuracy of glass modeling based on MD simulations largely depends on the reliability of the underlying empirical forcefield, i.e., the two-body (and sometimes three-body or more) interatomic potential [3,5]. Although *ab initio* molecular dynamics (AIMD) can, in theory, overcome these limitations, the high computational cost of this technique renders challenging glass simulations—which typically require large systems for statistical averaging and long timescales to slowly quench a melt down to the glassy state [3,6,7]. The development of new, improved empirical forcefields presently represents a bottleneck in glass modeling [8–10].

Empirical forcefields are typically based on functionals that depend on several parameters (e.g., partial atomic charges, etc.), which need to be properly optimized in order to minimize a given cost function [11,12]. One option is to define the cost function in terms of the difference between the structure or properties of the simulated system and available experimental data. However, such an optimization method may not yield a realistic forcefield in the case of glassy materials, since simulated and experimental glasses are prepared with significantly different cooling rate and, hence, their direct comparison may not be meaningful [6,13]. Although this problem can be partially overcome by conducting the optimization based on crystals rather than glasses, crystal-based potentials do not always properly describe the structure and properties of

disordered, out-of-equilibrium glasses [12]. Alternatively, empirical forcefield can be parameterized based on AIMD simulations [9,14,15]. However, directly optimizing the forcefield in order to match with the interatomic forces or energy derived from AIMD sometimes results in unrealistic structures for the simulated glasses [9,11,16]. Recently, Kob, Huang et al. proposed a new optimization scheme, wherein the optimization cost function is defined based on the difference between the structure of a simulated liquid and that obtained by AIMD simulations in similar conditions [9,11,15]. However, such cost functions are very "rough," that is, they exhibit a large number of local minima (i.e., several sets of parameters yield similar, competitive results). This is a challenge as conventional gradient-based optimization methods (e.g., steepest descent or conjugate gradient) are highly inefficient to explore rough functions and are likely to yield a local minimum rather than the global one [17]. Due to this issue, conventional optimization methods are often biased, that is, their outcomes strongly depend on the starting point.

As an alternative route to conventional "intuition-based" forcefield parameterization, artificial intelligence and machine learning (ML) techniques have the potential to offer some efficient, non-biased optimization schemes [18,19]. To this end, several ML-based forcefields have been proposed [8,20,21]. However, although such forcefields can approach the accuracy of AIMD at a fraction of computing cost, their parametrization remains tedious and the complex form of the resulting forcefields render challenging their physical interpretation and their implementation [10,20–22]. For these reasons, ML-based forcefields have thus far mostly been limited to simple systems (e.g., comprising only one element at a time [10,23]), which does not yet offer a realistic path toward the simulation of complex multi-component glasses.

Here, we present a new less accurate, but more pragmatic approach to efficiently parametrize forcefields based on ML-based optimization. Our method is based on a predefined

empirical potential form, wherein the parameters are optimized vs. AIMD simulations by Gaussian Process Regression and Bayesian optimization. We illustrate our new method by taking the example of glassy silica (g-SiO_2), an archetypal model for complex silicate glasses. Our method yields a new interatomic forcefield for g-SiO_2 that offers an unprecedented agreement with *ab initio* simulations. We demonstrate that, compared with traditional optimization methods, our ML-based optimization scheme is more efficient and non-biased. Overall, this work provides a realistic pathway toward the accurate, yet computationally efficient simulation of non-equilibrium disordered materials.

This paper is organized as follows. First, Sec. 7.2 describes the technical details of the simulations and parameterization strategy. The application of our method to glassy silica is then presented in Sec. 7.3. We then discuss the advantage of our approach over conventional optimization methods in Sec. 7.4. Finally, some conclusions are given in Sec. 7.5.

7.2 Methods

7.2.1 Reference *ab initio* simulations

A "reference" structure of a liquid silica system is first prepared by Car-Parrinello molecular dynamics (CPMD) [24]. The simulated system comprises 38 SiO_2 units (114 atoms) in a periodic cubic simulation box of length 11.982 Å—in accordance with the experimental density of 2.2 g/cm^3 [25]. The electronic structure is described with the framework of density functional theory and the choice of pseudopotentials for silicon and oxygen, exchange and correlation functions, and the plane-wave cutoff (70 Ry) are based on previous CPMD simulations of glassy silica [9,15]. A timestep of 0.0725 fs and a fictitious electronic mass of 600 atomic units are used. An initial liquid configuration is first prepared by conducting a classical MD run at 3600 K using

the well-established van Beest–Kramer–van Santen (BKS) potential (see Sec. 7.2.2) [14]. The obtained configuration is then relaxed via CPMD at 3600 K for 3.5 ps at constant volume. Such duration is long enough considering the small relaxation time of the system at such elevated temperature. A subsequent dynamics of 16 ps is then used for statistical averaging and to compute the Si–Si, Si–O, and O–O partial pair distribution functions (PDFs) of the simulated liquid system. More details can be found in Ref. [9,15].

7.2.2 Classical molecular dynamics simulations

A new empirical forcefield for g-SiO_2 is then parameterized by conducting some classical MD simulations. The simulated system comprises 1000 SiO_2 units (3000 atoms) in a periodic cubic simulation box of length 35.661 Å, which corresponds to the experimental density of 2.2 g/cm^3 [25]. An initial configuration is first prepared by relaxing the system for 10 ps at 3600 K in the *NVT* ensemble. The partial PDFs of the simulated systems are then computed based on a subsequent *NVT* dynamics of 10 ps. A timestep of 1 fs is used for all simulations.

The interatomic potential energy between each pair of atom i, j is here described by adopting the Buckingham form [9,14]:

$$U_{ij} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + A_{ij}\exp(-\frac{r_{ij}}{\rho_{ij}}) - \frac{C_{ij}}{r_{ij}^6} + \frac{D_{ij}}{r_{ij}^{24}} \qquad \textbf{Eq. (7-1)}$$

where r_{ij} is the distance between each pair of atoms, q_i is the partial charge of each atom (q_O for oxygen, q_{Si} for silicon, so that $q_O = -q_{Si}/2$), ε_0 is the dielectric constant, and the parameters A_{ij}, ρ_{ij}, C_{ij}, and D_{ij} describe the short-range interactions. A cutoff of 8 Å is used for the short-range interactions. The long-range coulombic interactions are evaluated by damped shifted force (dsf) model [26] with a damping parameter of 0.25 and a cutoff of 8 Å. The last term serves as to add a

strong repulsion at short distance to prevent the "Buckingham catastrophe" [9]. Since this term only aims to prevent any atomic overlap, the D_{ij} parameters are not included in the present optimization and their value is fixed based on Ref. [9] (viz., D_{ij} = 113, 29, and 3423200 eV·Å^{24} for O–O, Si–O, and Si–Si interactions, respectively). Note that this Buckingham formulation is chosen as it typically provides a good description of ionocovalent systems and has been shown to offer an improved description of g-SiO_2 as compared to alternative forms (e.g., Morse formulation) [11].

7.2.3 Optimization cost function

In total, the parametrization of this potential (Eq. (7-1)) requires the optimization of 10 independent parameters, namely, the partial charge q_{Si} and the short-range parameters $\{A_{ij}, \rho_{ij}, C_{ij}\}$ for each of the three atomic pairs (Si–O, O–O, and Si–Si). This set of parameters is denoted Ξ thereafter. Following Kob and Huang *et al.*, we define the optimization cost function R_χ as follows [9,11,15]:

$$R_\chi = \sqrt{\frac{\chi^2_{SiO} + \chi^2_{OO} + \chi^2_{SiSi}}{3}} \qquad \textbf{Eq. (7-2)}$$

where the $\chi^2_{\alpha\beta}$ terms capture the level of agreement between the partial PDFs obtained by classical MD and AIMD [27]:

$$\chi^2_{\alpha\beta} = \frac{\sum_r [g^{AIMD}_{\alpha\beta}(r) - g^{MD}_{\alpha\beta}(r)]^2}{\sum_r [g^{AIMD}_{\alpha\beta}(r)]^2} \qquad \textbf{Eq. (7-3)}$$

where $g^{AIMD}_{\alpha\beta}(r)$ and $g^{MD}_{\alpha\beta}(r)$ are the partial PDFs for each pair of atoms *α–β*.

Although additional properties (e.g., energy, stiffness, etc.) could be included, we herein restrict the cost function to the difference between AIMD and MD partial PDFs. This choice is

motivated by the following facts. (i) Although other structural descriptors could be used to describe the structure of the simulated glasses, the PDF offers a convenient description of the short-range environment around each atom [9,15] and, hence, capture some important features of the atomic structure. (ii) We purposely exclude from the training set any properties of glassy SiO_2 (e.g., experimental density or stiffness) as such properties are not uniquely defined and depend on the cooling rate. (iii) Including some additional properties (i.e., besides the PDFs) in the cost function would raise the question of which weight to attribute to each property—which would render the parameterization of the forcefield biased to this arbitrary choice.

7.2.4 Forcefield optimization by machine learning

We now describe the ML-based optimization scheme used herein to parametrize the forcefield. An overview of the parametrization process is presented in Fig. 7-1. First, we create an initial dataset comprising some "known points," that is, the values of the cost function R_χ for select sets of parameters Ξ. Gaussian Process Regression (GPR) [28,29] is then used to interpolate the known points and assess the interpolation uncertainty over the entire parameter space. The Bayesian Optimization (BO) based on expected improvement (EI) method [28] is then used to predict an optimal set of parameters Ξ that offers the best "exploration vs. exploitation trade-off," that is, the best balance between (i) exploring the parameter space and reducing the model uncertainty and (ii) finding the global minimum of the cost function. The cost function $R_\chi\{\Xi\}$ associated with the set of parameters predicted by BO is subsequently calculated by conducting a classical MD simulation and comparing the structure of the simulated liquid with that of the reference AIMD configuration (see Sec. 7.2.3). This new datapoint $R_\chi\{\Xi\}$ is then added to the dataset. The new dataset is then used to refine the GPR-based interpolation and predict a new

optimal set of parameters by BO. This cycle is iteratively repeated until a satisfactory minimum in the cost function is obtained (i.e., after about 600 iterations in this work). Finally, the global minimum predicted by BO is further refined by conducting a conjugate gradient (CG) optimization [17].

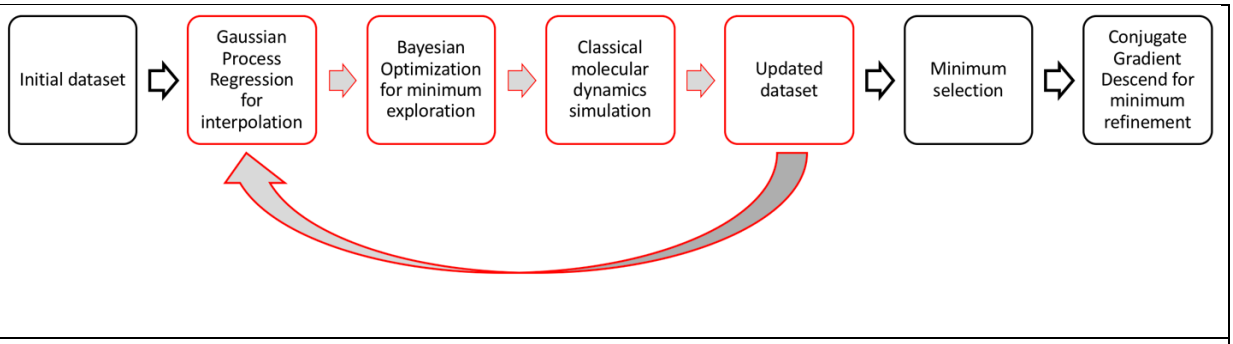


Figure 7-1: Flowchart diagram summarizing our parametrization strategy.

7.3 Results

7.3.1 New interatomic forcefield for glassy silica

We conduct the optimization of the forcefield while keeping the Si–Si interaction energy term as being zero (that is, $A_{SiSi} = C_{SiSi} = 0$ and $\rho_{SiSi} = 1$ Å). This choice is motivated by the fact that the original BKS potential does not comprise any Si–Si interaction energy terms, which suggests that the addition of these terms may not be necessary. In turn, decreasing the number of variable parameters allows us to increase the efficiency of the optimization. In addition, decreasing the complexity of the forcefield limits the risk of overfitting, which, in turn, is likely to increase the transferability of the new forcefield to new systems that are not considered during its training. The effect of the complexity of the forcefield (and of Si–Si terms) is further discussed in Ref. [30].

The forcefield parameters obtained after the BO and CG optimization steps are listed in Tab. 7-1. The performance of our forcefield (as quantified in terms of the final cost function R_χ) is

compared with that of select alternative potentials in Tab. 7-2. We find that our new ML forcefield yields an R_χ of 8.77%. This constitutes a significant improvement with respect to the well-established BKS potential (for which R_χ is about 17%) [14,31]. Our new potential is also found to be slightly better than the CHIK potential parameterized by Kob *et al.* [15]. This is not surprising as the CHIK potential was obtained based on the optimization of a slightly different cost function [15]. However, it is worth noting that our new potential exhibits a lower complexity than the CHIK parametrization (which comprises 3 extra parameters for the Si–Si interactions).

Table 7-1. Parameters of our new interatomic potential "ML" (see Eq. (7-1)). The partial charges are indicated as subscripts for each pair of atoms.

Atomic pairs	A (eV)	ρ (Å)	C (eV·Å^6)
$Si^{+1.955} – O^{-0.9775}$	20453.601	0.191735	93.496
$O^{-0.9775} – O^{-0.9775}$	1003.387	0.356855	81.491
$Si^{+1.955} – Si^{+1.955}$	0	1	0

In details, we find that the parameters of our ML forcefield are significantly different from those of the original BKS potential—which illustrates the roughness of the cost function. Interestingly, we find that our ML potential relies on a partial charge for Si atoms that is significantly smaller than that of the BKS potential (+1.955 vs. +2.4 for BKS). In turn, this value is close to that of the CHIK (+1.91 [15]) and Wang–Bauchy potential (+1.89 [12]). This suggests that "soft potentials" (i.e., which relies on lower partial charges) appear to consistently perform better than the stiffer ones, e.g., BKS. The cost function R_χ associated with each interatomic pair

(see Eq. (7-3)) is also provided in Tab. 7-2. Overall, we note that our ML potential consistently offers an improved description of the interatomic structural order for each pair of atoms.

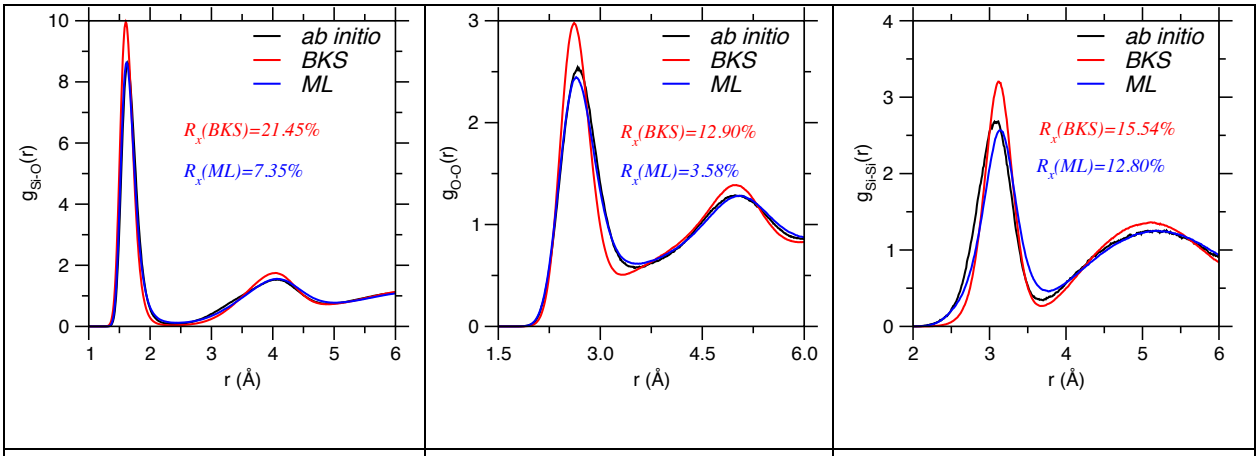
Table 7-2. Comparison of our new "ML" forcefield with select alternative classical potentials, namely, "BKS" [14] and "CHIK" [15].

Forcefield	R_χ^{SiO} (%)	R_χ^{OO} (%)	R_χ^{SiSi} (%)	Global R_χ (%)
ML	7.35	3.58	12.80	8.77 ± 0.25
BKS	21.45	12.90	15.54	17.01 ± 0.25
CHIK	12.29	6.09	11.76	10.43 ± 0.25

7.3.2 Partial pair distribution functions

We now further analyze the structure of the simulated SiO_2 liquid (i.e., at 3600 K). Fig. 7-2 shows the partial PDFs predicted by our new ML forcefield. The data are compared with the reference *ab initio* partial PDFs used for the training of the potential [15] as well as those predicted by the BKS potential [14]. Overall, we find that our ML forcefield offers an excellent agreement with AIMD simulations—although this is not surprising as our forcefield is specifically trained to match these data. Nevertheless, these results show that the Buckingham formulation adopted herein is appropriate for the SiO_2 system and further supports the ability of our optimization method to offer a robust parametrization. We note that the average Si–Si distance predicted by our potential is slightly shifted with respect to that obtained in AIMD simulations (see Fig. 7-2(c)). This may arise from a general limitation of the Buckingham formulation. Nevertheless, our ML forcefield offers a significant improvement with respect to the BKS potential, especially in the case of the Si–O and O–O partial PDFs (see also Tab 7-2). We note that our ML forcefield

systematically predicts some PDF peaks that are broader than those predicted by BKS, which suggests that our forcefield yields a slightly more disordered structure. This may be linked with the fact that our potential relies on lower partial charge values (i.e., softer Coulombic interactions).



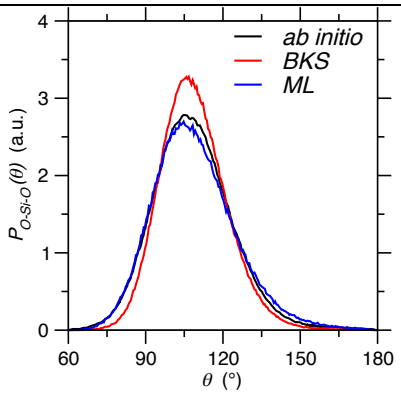
(a) **(b)** **(c)**

Figure 7-2: (a) Si–O, **(b)** O–O, and **(c)** Si–Si partial pair distribution functions (PDFs) in liquid silica (at $T = 3600$ K) predicted by our new "ML" forcefield and compared with the *ab initio* reference [15]. The partial PDFs predicted by the BKS potential are added for comparison [14].

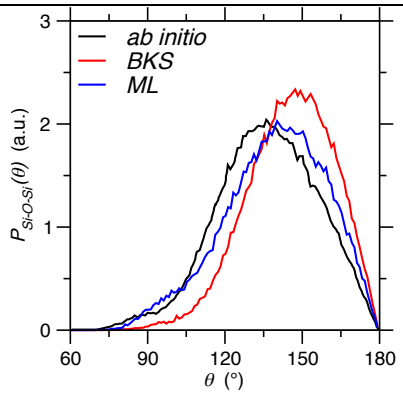
7.3.3 Partial bond angle distributions

We now focus on the angular environment around each atom. To this end, Fig. 7-3 shows the O–Si–O and Si–O–Si partial bond angle distributions (PBADs) predicted by our ML forcefield for the liquid silica system (at $T = 3600$ K). The data are compared with those obtained by *ab initio* simulations [15] and predicted by the BKS potential [14]. Overall, we observe that the PBADs predicted by our ML forcefield are in very good agreement with those obtained by *ab initio* simulations—with a significant improvement with respect to the BKS potential. This is significant as the PBADs are not explicitly included in the cost function used herein and such 3-body

correlations are not fully encoded in 2-body correlations (i.e., as captured by the partial PDFs). As such, these results offer a strong a posteriori validation of the performance of our new ML forcefield.



(a)



(b)

Figure 7-3: (a) O–Si–O and **(b)** Si–O–Si partial bond angle distributions (PBADs) in liquid silica (at $T = 3600$ K) predicted by our new "ML" forcefield and compared with the *ab initio* reference [15]. The PBADs predicted by the BKS potential are added for comparison [14].

As expected, our forcefield yields a tetrahedral environment for Si atoms (with an average O–Si–O angle of about 109°). However, we note that the O–Si–O PBAD predicted by our ML forcefield is broader than that obtained with BKS, which suggests that our potential yields a slightly more disordered angular environment around Si atoms. Again, this may be linked with the fact that our potential relies on lower fictive charges than BKS (see Sec. 7.3.2). In contrast, we observe that our forcefield slightly overestimates the value of Si–O–Si angle with respect to AIMD simulations. This is likely linked with the fact that our potential overestimates the Si–Si average distance (see Sec. 7.3.2), which appears to be a general limitation of the two-body Buckingham

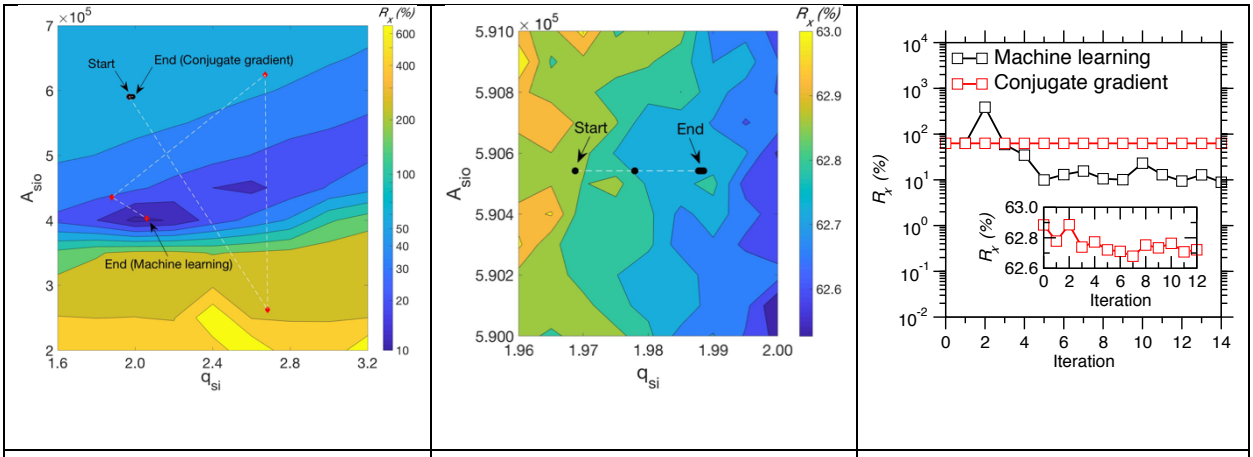
formulation. Nevertheless, the Si–O–Si PBAD yielded by our forcefield is significantly improved with that obtained by BKS (which tends to largely overestimate the average Si–O–Si angle).

7.4 Discussion

7.4.1 Comparison between gradient-based and machine-learning-based optimization

We now discuss the performance of our ML-based optimization method by comparing its ability to identify the global minimum of the cost function with that of the conjugate gradient method. Here, for illustrative purposes, only two parameters (q_{Si} and A_{SiO}) are optimized in both cases, while the other 8 forcefield parameters are kept fixed and equal to those found in the original BKS potential [14]. As shown in Fig. 7-4(a), the cost function R_χ shows a very rough dependence on the forcefield parameters—wherein the level of roughness appears to increase when upon zooming on the fine details of the landscape (see Fig. 7-4(b)). The pathways explored (starting from the same initial point) upon the ML-based and CG-based optimizations in the (q_{Si}, A_{SiO}) space is shown in Fig. 7-4(a). We observe that the ML-based optimization quickly converges toward the global minimum of the cost function after only 5 iterations, after which the cost function R_χ shows a plateau around 10% (see Fig. 7-4(c)). This illustrates the efficiency of our optimization technique. In contrast with our ML optimization method, the CG optimization quickly gets "stuck" in a local minimum of the cost function (see Fig. 7-4(c)) and does not succeed at identifying the global minimum. This highlights the fact that traditional gradient-based optimization methods are not appropriate in the case of such high-roughness function and, hence, are highly biased based on the chosen starting point. Although the efficiency of the CG method could certainly be improved by adjusting some parameters (e.g., the learning rate and step length [17]), such fine-tuning

necessarily requires some level of intuition or trial-and-error optimization, which is a clear advantage of the present ML approach.



(a) **(b)** **(c)**

Figure 7-4: Comparison between machine learning and conjugate gradient optimization. Only the partial charge of the Si atoms q_{Si} and the parameter A_{SiO} are here optimized, while the other 8 forcefield parameters are kept fixed. **(a)** Contour plot showing the cost function R_χ as a function of q_{Si} and A_{SiO}. The red and black circles indicate the path explored upon machine learning and conjugate gradient optimization, respectively. Panel **(b)** is a zoom of the data presented in panel (a) to better observe the path explored upon conjugate gradient optimization. **(c)** Evolution of the cost function R_χ during the machine learning and conjugate gradient optimizations. The inset is a zoom on data obtained in the case of conjugate gradient optimization.

7.4.2 Lessons from the BKS potential

It worth further focusing on the BKS potential [14] to establish some general conclusions regarding the development of interatomic forcefields for glassy materials. The well-established BKS potential was parameterized by sequentially optimizing the O–O and Si–O energy terms so

as to an isolated SiO_4 cluster (saturated by 4 H atoms) matches with *ab initio* simulations. Si–Si energy terms were forced to be zero. In addition, the experimental elastic constants of silica were used to discriminate several competing sets of parameters. This suggests that the BKS potential is specifically trained to offer an excellent description of the interatomic potential in the vicinity of its equilibrium position. In details, the position of the minima of each energy terms is encoded in the geometry of the isolated SiO_4 cluster (i.e., the average interatomic distances), while the curvature of the potential energy at the vicinity of the equilibrium position is encoded in the elastic constant. Nevertheless, as detailed in Sec. 7.3, this optimization scheme tends to overestimate the radial and angular order around Si atoms (see Figs. 7-2 and 7-3). This suggests that optimization schemes placing a strong emphasis on describing the shape of the forcefield in the very close vicinity of the equilibrium position may not be appropriate to describe the disordered structure of glasses, which are intrinsically out-of-equilibrium and wherein the atoms are not exactly located at their minimum-energy positions. For instance, the degree of asymmetry of the forcefield is likely to play a key role in governing the structure of disordered materials and may not be efficiently trained by considering only equilibrium structures (e.g., crystals or isolated clusters). This suggests that parametrization methods based on liquid structures (as the present one) may be more appropriate to develop new improved forcefields for complex glasses.

7.5 Conclusions

Overall, this study establishes a general and versatile framework to accelerate the parametrization of new, improved empirical forcefields for disordered materials. As shown herein with the example of silica, our method makes it possible to quickly reoptimize previous well-established potentials (e.g., the BKS forcefield). By using as a reference some liquid structures

prepared by *ab initio* molecular dynamics simulations, our parametrization scheme is better suited for glass modeling than alternative methods based on equilibrium crystal or isolated atomic clusters. Importantly, the use of machine learning rather than alternative traditional optimization methods (e.g., conjugate gradient) (i) drastically improves the efficiency of the parametrization procedure, (ii) suppresses the risk of bias resulting from arbitrary choices regarding the starting point of the optimization, and (iii) significantly reduces the role played by "personal intuition" during the parametrization. As a key advantage over alternative conventional method, the present ML-based parametrization method is highly scalable and, hence, can be used to parametrize multi-component systems (i.e., many forcefield parameters can be optimized simultaneously). Overall, this work establishes an efficient, pragmatic method to develop new improved forcefields for the simulation of complex "real-world" materials—which addresses an immediate concern since more accurate ML-based forcefields that do rely on a predefined functional are unlikely to be available for complex multi-component systems in the near future.

7.6 References

[1] P.Y. Huang, S. Kurasch, J.S. Alden, A. Shekhawat, A.A. Alemi, P.L. McEuen, J.P. Sethna, U. Kaiser, D.A. Muller, Imaging Atomic Rearrangements in Two-Dimensional Silica Glass: Watching Silica's Dance, Science. 342 (2013) 224–227. https://doi.org/10.1126/science.1242248.

[2] L. Huang, J. Kieffer, Challenges in Modeling Mixed Ionic-Covalent Glass Formers, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials, Springer International Publishing, 2015: pp. 87–112. https://doi.org/10.1007/978-3-319-15675-0_4.

[3] J. Du, Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials, Springer International Publishing, 2015: pp. 157–180.

[4] M. Bauchy, Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: A review, Computational Materials Science. 159 (2019) 95–102. https://doi.org/10.1016/j.commatsci.2018.12.004.

[5] Y. Yu, B. Wang, M. Wang, G. Sant, M. Bauchy, Revisiting silica with ReaxFF: Towards improved predictions of glass structure and properties via reactive molecular dynamics, Journal of Non-Crystalline Solids. 443 (2016) 148–154. https://doi.org/10.1016/j.jnoncrysol.2016.03.026.

[6] X. Li, W. Song, K. Yang, N.M.A. Krishnan, B. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Balonis, M. Bauchy, Cooling rate effects in sodium silicate glasses: Bridging the gap between molecular dynamics simulations and experiments, The Journal of Chemical Physics. 147 (2017) 074501. https://doi.org/10.1063/1.4998611.

[7] P. Ganster, M. Benoit, J.-M. Delaye, W. Kob, Structural and vibrational properties of a calcium aluminosilicate glass: classical force-fields vs. first-principles, Molecular Simulation. 33 (2007) 1093–1103.

[8] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (2016) 170901. https://doi.org/10.1063/1.4966192.

[9] A. Carré, S. Ispas, J. Horbach, W. Kob, Developing empirical potentials from ab initio simulations: The case of amorphous silica, Computational Materials Science. 124 (2016) 323–334. https://doi.org/10.1016/j.commatsci.2016.07.041.

[10] A.P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, Physical Review X. 8 (2018). https://doi.org/10.1103/PhysRevX.8.041048.

[11] S. Sundararaman, L. Huang, S. Ispas, W. Kob, New optimization scheme to obtain interaction potentials for oxide glasses, J. Chem. Phys. 148 (2018) 194504. https://doi.org/10.1063/1.5023707.

[12] M. Wang, N.M.A. Krishnan, B. Wang, M.M. Smedskjaer, J.C. Mauro, M. Bauchy, A new transferable interatomic potential for molecular dynamics simulations of borosilicate glasses, Journal of Non-Crystalline Solids. 498 (2018) 294–304. https://doi.org/10.1016/j.jnoncrysol.2018.04.063.

[13] J.M.D. Lane, Cooling rate and stress relaxation in silica melts and glasses via microsecond molecular dynamics, Phys. Rev. E. 92 (2015) 012320. https://doi.org/10.1103/PhysRevE.92.012320.

[14] B.W.H. van Beest, G.J. Kramer, R.A. van Santen, Force fields for silicas and aluminophosphates based on *ab initio* calculations, Physical Review Letters. 64 (1990) 1955–1958. https://doi.org/10.1103/PhysRevLett.64.1955.

[15] A. Carré, J. Horbach, S. Ispas, W. Kob, New fitting scheme to obtain effective potential from Car-Parrinello molecular-dynamics simulations: Application to silica, EPL. 82 (2008) 17001. https://doi.org/10.1209/0295-5075/82/17001.

[16] F. Ercolessi, J.B. Adams, Interatomic Potentials from First-Principles Calculations: The Force-Matching Method, EPL. 26 (1994) 583. https://doi.org/10.1209/0295-5075/26/8/005.

[17] J.R. Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, 1994. https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

[18] J.E. Gubernatis, T. Lookman, Machine learning in materials design and discovery: Examples from the present and suggestions for the future, Physical Review Materials. 2 (2018). https://doi.org/10.1103/PhysRevMaterials.2.120301.

[19] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, Npj Computational Materials. 3 (2017) 54. https://doi.org/10.1038/s41524-017-0056-5.

[20] T.D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, A universal strategy for the creation of machine learning-based atomistic force fields, Npj Computational Materials. 3 (2017) 37. https://doi.org/10.1038/s41524-017-0042-y.

[21] Y. Li, H. Li, F.C. Pickard, B. Narayanan, F.G. Sen, M.K.Y. Chan, S.K.R.S. Sankaranarayanan, B.R. Brooks, B. Roux, Machine Learning Force Field Parameters from Ab Initio Data, J. Chem. Theory Comput. 13 (2017) 4492–4503. https://doi.org/10.1021/acs.jctc.7b00521.

[22] M. Hellström, J. Behler, Neural Network Potentials in Materials Modeling, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling, Springer International Publishing, Cham, 2018: pp. 1–20. https://doi.org/10.1007/978-3-319-42913-7_56-1.

[23] V.L. Deringer, G. Csányi, Machine learning based interatomic potential for amorphous carbon, Phys. Rev. B. 95 (2017) 094203. https://doi.org/10.1103/PhysRevB.95.094203.

[24] R. Car, M. Parrinello, Unified Approach for Molecular Dynamics and Density-Functional Theory, Physical Review Letters. 55 (1985) 2471–2474. https://doi.org/10.1103/PhysRevLett.55.2471.

[25] N.P. Bansal, R.H. Doremus, Handbook of Glass Properties, Elsevier, 2013.

[26] C.J. Fennell, J.D. Gezelter, Is the Ewald summation still necessary? Pairwise alternatives to the accepted standard for long-range electrostatics, The Journal of Chemical Physics. 124 (2006) 234104. https://doi.org/10.1063/1.2206581.

[27] A.C. Wright, The comparison of molecular dynamics simulations with diffraction experiments, Journal of Non-Crystalline Solids. 159 (1993) 264–268. https://doi.org/10.1016/0022-3093(93)90232-M.

[28] P.I. Frazier, J. Wang, Bayesian Optimization for Materials Design, in: Information Science for Materials Discovery and Design, Springer, Cham, 2016: pp. 45–75. https://doi.org/10.1007/978-3-319-23871-5_3.

[29] C.E. Rasmussen, C.K.I. Williams, Gaussian processes for machine learning, 3. print, MIT Press, Cambridge, Mass., 2008.

[30] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Balance between Accuracy and Simplicity in Empirical Forcefields for Glass Modeling: Insights from Machine Learning, Journal of Non-Crystalline Solids. (2019).

[31] B. Wang, Y. Yu, Y.J. Lee, M. Bauchy, Intrinsic Nano-Ductility of Glasses: The Critical Role of Composition, Front. Mater. 2 (2015) 11. https://doi.org/10.3389/fmats.2015.00011.

Chapter 8. Balance between Accuracy and Simplicity in Empirical Forcefields for Glass Modeling: Insights from Machine Learning

8.1 Introduction

The development of accurate, yet transferable empirical forcefields is key to model multicomponent glasses by molecular dynamics (MD) or Monte Carlo simulations [1,2]. To this end, several forms of empirical potentials are available, ranging from very simple (e.g., Lennard Jones potential) to very complex (e.g., ReaxFF potential [3–5]). The degree of complexity of empirical forcefields mostly depends on the number of parameters that need to be parameterized, which can range from 2 (for Lennard Jones potentials) to hundreds (for ReaxFF) of parameters for pairs of elements. As such, the parameterization of a new forcefield typically follows two steps: (i) selecting an appropriate analytical form and degree of complexity and (ii) optimizing the value of the forcefield parameters [2,6,7].

The second step has been extensively addressed, as several methods have been proposed to optimize the parameters of a given forcefield formulation to properly describe the structure and properties of a given system. The parameterization of a forcefield can usually be described as an optimization problem, wherein a given cost function needs to be minimized. On the one hand, the cost function can be defined based on the difference between the structure or properties of simulated and experimental glasses. However, this approach can be problematic as the cooling rates used in MD simulations and experiments are dramatically different, which renders challenging a meaningful comparison between simulated and experimental glasses [8–10]. On the other hand, for a given system, the cost function can be defined based on the difference between the outcomes of classical and *ab initio* molecular dynamics (AIMD) simulations [11–13]. Kob, Huang *et al.* have recently proposed a new forcefield parameterization strategy that consists in

defining the cost function in terms of the difference between the pair distribution function of a liquid simulated by AIMD and classical MD (i.e., as predicted by the forcefield that is to be trained) [11,14,15]. However, this cost function is very "rough," that is, it exhibits many local minima—i.e., the parametrization can yield several forcefields with different parameters, yet competitive accuracy [11,15]. As such, the outcome of the parameterization strongly depends on the starting point that is used [16]—so that the parameterization of the potential requires some level of "intuition."

In contrast, the first step of forcefield parameterization (i.e., selecting an appropriate degree of complexity) has received very little attention and often remains entirely based on "intuition" or "previous experience." However, selecting the right level of complexity is key to obtain accurate, yet transferable potentials. In details, forcefields that are too simple may not properly describe complex systems—for instance, Lennard Jones only rely on two parameters and, hence, are usually unable to properly predict at the same time the molar volume, molar energy, and stiffness of even simple systems (e.g., perfect gas). In contrast, forcefields that are too complex may offer an extremely accurate description of a targeted system, but offer very poor predictions when applied to systems that were not explicitly accounted for during the training of the forcefield (i.e., low transferability to new systems). In general, this competition between accuracy, transferability, and simplicity is known as a balance between "underfitting" and "overfitting."

Here, we report a new forcefield parametrization method that is based on machine learning (ML), which aims to reduce/suppress the need for intuition when (i) selecting the appropriate level of complexity for a forcefield and (ii) optimizing the value of the forcefield parameters. To illustrate this method, we take the examples of glassy silica as a system and of a Buckingham formulation for the forcefield. Our method allows us to quickly and robustly identify some optimal

forcefield parameters for different degrees of forcefield complexity and, based on these results, to identify the optimal balance between forcefield accuracy and simplicity. Overall, our method could greatly accelerate the development of new accurate, yet transferable forcefield for the modeling of silicate glasses.

This paper is organized as follows. First, Sec. 8.2 describes the forcefield formulation (and complexity thereof) that is adopted herein and offers a detailed description of our ML-based parameterization strategy. We then investigate the influence of the forcefield complexity in Sec. 8.3 and 8.4. Finally, some conclusions are given in Sec. 8.5.

8.2 Methods

8.2.1 Empirical forcefields of different complexity

Glassy silica (g-SiO_2) is an archetypal ionocovalent system—whose interatomic potential energy can be well described by the Buckingham form relying only on two-body interactions between each pair of atom i, j [6,11,12]:

$$U_{ij} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + A_{ij}\exp\left(-\frac{r_{ij}}{\rho_{ij}}\right) - \frac{C_{ij}}{r_{ij}^6} + \frac{D_{ij}}{r_{ij}^{24}} \qquad \textbf{Eq. (8-1)}$$

where r_{ij} is the distance between each pair of atoms, q_i are the partial charges of each atom (q_{O} for oxygen, q_{Si} for silicon, so that $q_{\mathrm{O}} = -q_{\mathrm{Si}}/2$), ε_0 is the dielectric constant, and the parameters A_{ij}, ρ_{ij}, C_{ij}, and D_{ij} describe the short-range interactions. A cutoff of 8 Å is here consistently used for the short-range interactions. The long-range coulombic interactions are evaluated by damped shifted force (dsf) model [17] with a damping parameter of 0.25 and a cutoff of 8 Å. Here, the last term is added as a strong repulsion at short-distance to avoid the "Buckingham catastrophe" [11], wherein the D_{ij} parameter is fixed to prevent any atomic overlap based on Ref. [11] (*viz.*, D_{ij} = 113, 29, and 3423200 eV·Å^{24} for O–O, Si–O, and Si–Si interactions, respectively). In total, 10

independent parameters need to be parameterized for this forcefield formulation (Eq. (8-1)), namely, the partial charge q_{Si} and the short-range parameters $\{A_{ij}, \rho_{ij}, C_{ij}\}$ for each of the three atomic pairs (Si–O, O–O, and Si–Si). This set of parameters is denoted Ξ thereafter.

In the present case, the degree of complexity of this forcefield can be quantified by the number of parameters that are non-zero (out of the 10 independent parameters). For instance, although they are both based on the same Buckingham formulation, the well-established van Beest–Kramer–van Santen (BKS) [12] potential does not comprise any Si–Si energy terms, whereas such terms are present within the Carré–Horbach–Ipsas–Kob (CHIK) potential [11]. Here, to assess the influence of the potential complexity, we parameterize via a novel ML approach three potentials featuring an increasing level of complexity, namely (i) ML-SiO, wherein only Si–O interaction energy terms are considered (i.e., 4 non-zero parameters in Ξ), (ii) ML, wherein only Si–O and O–O interaction energy terms are considered (i.e., 7 non-zero parameters in Ξ), and (iii) ML-ALL, wherein all the Si–O, O–O, and Si–Si interaction energy terms are considered (i.e., 10 non-zero parameters in Ξ).

8.2.2 Forcefield parameterization from *ab initio* simulation

Following Kob and Huang *et al.*, the determination of the optimal parameters Ξ is conducted by minimizing the difference between the outcomes of classical MD and AIMD while simulating an equilibrium silica liquid [11,14,15]. To this end, we define the cost function R_χ as follows:

$$R_\chi = \sqrt{\frac{\chi^2_{\mathrm{SiO}} + \chi^2_{\mathrm{OO}} + \chi^2_{\mathrm{SiSi}}}{3}} \qquad \textbf{Eq. (8-2)}$$

where the $\chi^2_{\alpha\beta}$ terms capture the level of agreement between the partial pair distribution functions (PDFs) obtained by classical MD and AIMD [18]:

$$\chi^2_{\alpha\beta} = \frac{\sum_r [g^{\mathrm{AIMD}}_{\alpha\beta}(r) - g^{\mathrm{MD}}_{\alpha\beta}(r)]^2}{\sum_r [g^{\mathrm{AIMD}}_{\alpha\beta}(r)]^2} \qquad \textbf{Eq. (8-3)}$$

where $g^{\mathrm{AIMD}}_{\alpha\beta}(r)$ and $g^{\mathrm{MD}}_{\alpha\beta}(r)$ are the partial PDFs for each pair of atoms α–β. Note that, among potential alternative structural metric describing the structure of the simulated glasses or liquids, the PDF offers a convenient description of the short-range environment around each atom [11,14,19]. We purposely exclude from the training set any of the properties of glassy SiO_2 (e.g., experimental density or stiffness) as such properties are not uniquely defined and depend on the cooling rate. This training scheme is motivated by the fact that Buckingham-type potentials have been shown to properly describe (i) the temperature-dependence of glass and liquid properties and (ii) the dependence of glass properties on the cooling rate (see Refs. [8–10]), so that training the system for a fixed temperature should yield a good description of its behavior as a function of temperature, including in the glassy state. A similar approach was used in Refs. [11,14]. The technical details of MD and AIMD simulations are provided below.

(i) Reference AIMD simulations

The "reference" liquid silica structure is prepared by Car-Parrinello molecular dynamics (CPMD) [20]. 38 SiO_2 units (114 atoms) are placed within a periodic cubic simulation box of length 11.982 Å to match the experimental density of 2.2 g/cm^3 [21]. The electronic structure of the atoms is described within the framework of density functional theory. The choice of pseudopotentials for each atom-type, exchange and correlation functions, and the plane-wave cutoff (70 Ry) are based on previous CPMD simulations of glassy silica [11,14]. A timestep of 0.0725 fs and a fictitious electronic mass of 600 atomic units are used. A liquid configuration obtained by classical MD simulation at 3600 K using the well-established BKS potential is used

an initial configuration (see Sec. 8.2.1) [12]. This configuration is then relaxed via CPMD at 3600 K and constant volume for 3.5 ps—which duration is long enough due to the small relaxation time of the system at such elevated temperature. A subsequent dynamics of 16 ps is then used for statistical averaging and to compute the Si–Si, Si–O, and O–O PDFs of the simulated liquid system. Note that, although certain properties strongly depend on the system size (e.g., ring size distribution, vibrational properties, transport properties, etc.), partial PDFs have been shown to be fairly unaffected by the system size (as long as it is larger than 100 atoms, see Refs. [13,22]). More details on the CPMD simulations can be found in Ref. [11,14].

(ii) Classical MD simulation

For each set of forcefield parameters Ξ considered herein, we conduct a classical MD simulation of the same liquid silica system. The simulated system comprises 1000 SiO_2 units (3000 atoms) placed in a periodic cubic simulation box of length 35.661 Å—in accordance with the experimental density of 2.2 g/cm^3 [21]. The configuration is first fully relaxed for 10 ps at 3600 K in the *NVT* ensemble. The partial PDFs of the simulated systems are then computed based on statistical averaging in a subsequent *NVT* dynamics of 10 ps. A timestep of 1 fs is consistently used for all simulations.

8.2.3 Machine learning forcefield optimization

We now introduce the ML-based optimization scheme that is used to minimize the cost function and, thereby, parametrize the three forcefields considered herein (i.e., with different level of complexity). Fig. 8-1 shows an overview of the parametrization process. First, we construct an initial dataset containing some "known points," that is, the values of the cost function R_χ for select sets of parameters Ξ. This dataset serves as a training set for the machine learning algorithm, which

is able to "learn by example" the relationship between the parameters Ξ and the cost function R_χ. To this end, we use Gaussian Process Regression (GPR) [23,24] to interpolate the known points (and assess the uncertainty of the interpolation) over the entire parameter space. The Bayesian Optimization (BO) method [23] is then used to analyze the interpolated function and its uncertainty in order to predict an optimal set of parameters Ξ—which offers the best "exploration vs. exploitation trade-off", that is, the best balance between (i) exploring the parameter space and reducing the model uncertainty and (ii) exploiting the present apparent minimum and finding the global minimum of the cost function. The "true" cost function $R_\chi\{\Xi\}$ associated with the set of parameters predicted by BO is subsequently calculated by conducting a classical MD simulation and comparing the simulated structure with the reference AIMD configuration (see Sec. 8.2.2). This new datapoint $R_\chi\{\Xi\}$ is then added to the dataset. The new dataset is then used to refine the GPR-based interpolation and predict a new optimal set of parameters by BO. This cycle is iteratively repeated until a satisfactory minimum in the cost function is obtained, that is, when R_χ does not decrease any further. Finally, the global minimum predicted by BO is further refined by conducting a conjugate gradient (CG) optimization [16]. Each of these steps is further described in the following.

(i) Initial dataset

As a starting point for our optimization method, we construct an initial dataset, which contains as inputs a selection of potential parameters Ξ and as outputs the associated cost function R_χ. Each of these datapoints is obtained by an independent MD simulation (see Secs. 8.2.2). This initial dataset offers an ensemble of known values for the cost function in 10-dimensional parameter space (i.e., for the 10 components in Ξ), which is used as a starting point for the iterative interpolation/exploration process described in the following. These initial values of Ξ are chosen

so as to uniformly span the targeted range of parameters (chosen based on previously available forcefield). In the case of wide target ranges, we divide the target range into several small pitches for fast exploration. The initial dataset comprises about 1000 known points, which corresponds to a minuscule fraction of the parameter space. For instance, with 10 independent parameters, considering only two values for each parameter would yield $2^{10} = 1024$ possible combinations. Each known point is obtained by conducting an MD simulation that takes about 1 minute of computation using 16 CPU cores. Overall, it takes about 17 hours to establish the initial dataset.

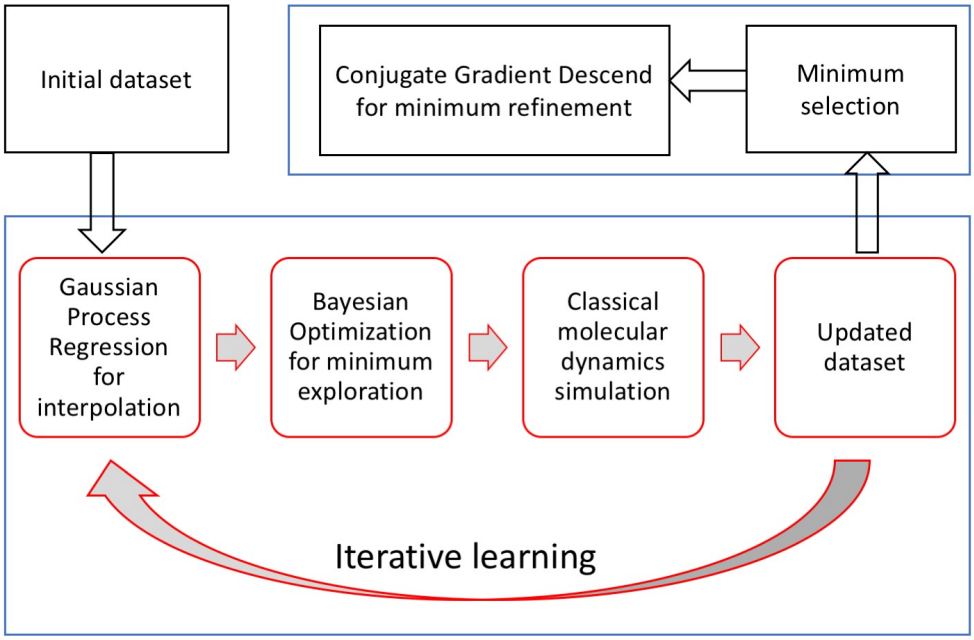
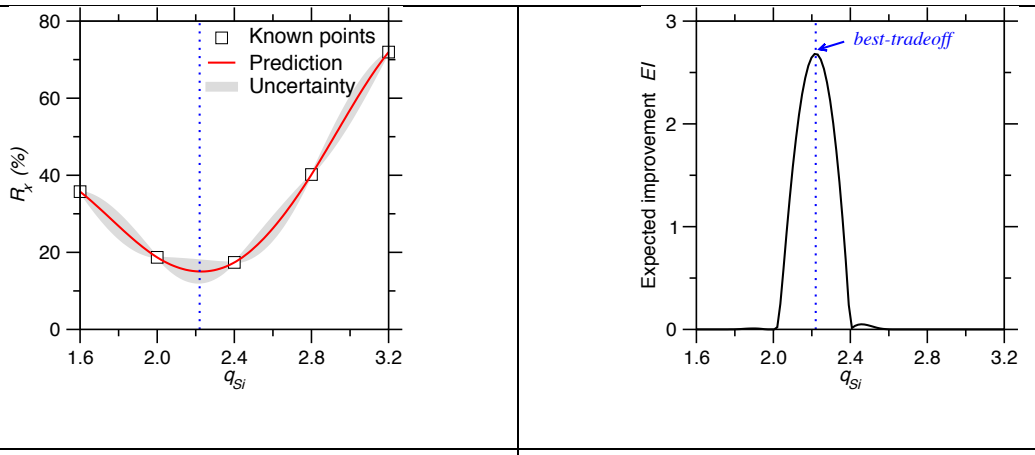


Figure 8-1: Flow-chart of the machine-learning-based parametrization strategy.

(ii) Interpolation by Gaussian Process Regression

The basic principle of GPR is to infer the (Gaussian-type) probability distribution of the values of the function that is interpolated based on a set of known points [23,24]. The interpolation process follows the following expression:



(a) **(b)**

Figure 8-2: Illustration of the Bayesian optimization approach used herein. Only the partial charge of the Si atoms q_{Si} is here optimized, while the other 9 forcefield parameters are kept fixed. **(a)** Interpolation of the cost function (R_χ, see Eq. (8-2)) offered by Gaussian Process Regression (red line) as a function of the q_{Si}. The prediction is based on an initial training set comprising 5 datapoints (i.e., known points, black symbols). The grey area indicates the uncertainty (95% confidence interval) of the prediction. **(b)** Expected Improvement (EI) function yielded by the Bayesian optimization method, which predicts the set of parameters (here, q_{Si}) that offers the best tradeoff between "exploration" (i.e., minimizing the uncertainty of the model presented in panel (a)) and "exploitation" (i.e., minimizing the cost function R_χ).

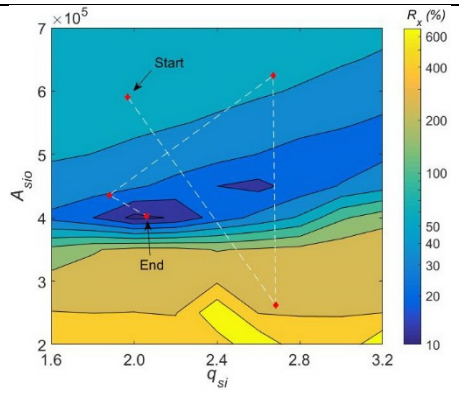
$$P\left(R_\chi(\Xi^*)\middle|\left\{R_\chi(\Xi_{\mathrm{known}})\right\}\right) \Leftarrow$$

$$\begin{bmatrix} R_\chi(\Xi_1) \\ \vdots \\ R_\chi(\Xi_n) \\ R_\chi(\Xi^*) \end{bmatrix} \sim \mathrm{Normal}\left(\begin{bmatrix} \mu_0(\Xi_1) \\ \vdots \\ \mu_0(\Xi_n) \\ \mu_0(\Xi^*) \end{bmatrix} , \begin{bmatrix} \Sigma_0(\Xi_1,\Xi_1) & \cdots & \Sigma_0(\Xi_1,\Xi_n) & \Sigma_0(\Xi_1,\Xi^*) \\ \vdots & \ddots & \vdots & \vdots \\ \Sigma_0(\Xi_n,\Xi_1) & \cdots & \Sigma_0(\Xi_n,\Xi_n) & \Sigma_0(\Xi_n,\Xi^*) \\ \Sigma_0(\Xi^*,\Xi_1) & \cdots & \Sigma_0(\Xi^*,\Xi_n) & \Sigma_0(\Xi^*,\Xi^*) \end{bmatrix} \right) \qquad \textbf{Eq. (8-4)}$$

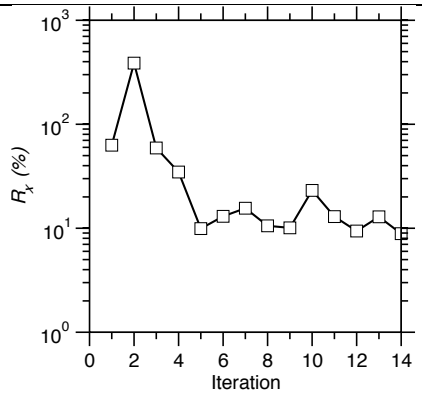
where $P\left(R_\chi(\Xi^*)\middle|\left\{R_\chi(\Xi_{\mathrm{known}})\right\}\right)$ is the conditional probability of the value of the cost function R_χ for a given set of parameters Ξ^* given the dataset of all the known points

$\{R_\chi(\Xi_1),\ \ R_\chi(\Xi_2),\ \ \cdots\ \ R_\chi(\Xi_n)\}$, as denoted as $\{R_\chi(\Xi_{\text{known}})\}$. The conditional probability of $R_\chi(\Xi^*)$ is calculated using multivariate Gaussian distribution [25], where $\mu_0(\cdot)$ is the mean operation and $\Sigma_0(\cdot)$ is the covariance operation. There are many possible choices for the function-type of $\mu_0(\cdot)$ and $\Sigma_0(\cdot)$ and most can offer a reasonable extrapolation in the framework of multivariate Gaussian distribution [25]. Here, we adopt the Matern-type kernel for $\mu_0(\cdot)$ and $\Sigma_0(\cdot)$ [24,25]. In addition, to add some white-noise background during the interpolation [23], we also checked the intrinsic uncertainty of the cost function values yielded by the MD simulations by conducting a series of 10 independent MD simulations while keeping the same set of parameters Ξ and calculating the standard deviation of the associated cost functions R_χ. We find that the computed cost function values have a relative uncertainty of about 2% when $R_\chi < 100\%$ (i.e., for realistic forcefields) and can increase up to 10% for higher values of R_χ (i.e., for fairly unrealistic forcefields). This level of noise is not expected to significantly affect the shape of interpolation around the minimum positions of the cost function R_χ.

Fig. 8-2(a) shows an example of the outcome of a GPR-based interpolation. For illustration purposes, only the partial charge of the Si atoms q_{Si} is here optimized, while the other 9 forcefield parameters are kept fixed and equal to those found in the original BKS potential [12]. A dataset comprising the values of the cost function R_χ for 5 values of q_{Si} ranging from 1.6-to-3.2 is first constructed. The interpolated function and the uncertainty thereof (95% confidence interval) predicted by GPR is shown in Fig. 8-2(a). As expected, we observe that the interpolated function exhibits a minimum with respect to q_{Si} (note that the q_{Si} value used in the BKS potential is 2.4). Unsurprisingly, the uncertainty of the prediction is low at the vicinity of the known points and increases in between them.



(a)



(b)

Figure 8-3: Illustration of the iterative optimization approach used herein. Only the partial charge of the Si atoms q_{Si} and the parameter A_{SiO} are here optimized, while the other 8 forcefield parameters are kept fixed. **(a)** Contour plot showing the cost function R_χ as a function of q_{Si} and A_{SiO}. The white dashed line indicates the path explored by the Bayesian optimization method until the global minimum in the cost function R_χ is identified. **(b)** Evolution of the cost function R_χ of the best-tradeoff position predicted by the Bayesian optimization during the optimization process.

(iii) Minimum exploration by Bayesian optimization

Based on the interpolated function $R_\chi(\Xi)$ and uncertainty $\sigma(\Xi)$ thereof predicted by GPR, the BO method is used to determine the next optimal set of parameters Ξ to try based on an acquisition function that depends on $R_\chi(\Xi)$ and $\sigma(\Xi)$. Here, we adopt the expected improvement (EI) function, which is commonly used as acquisition function [23]:

$$EI(\Xi) = \begin{cases} \left[R_\chi\left(\hat{\Xi}\right) - R_\chi(\Xi)\right]\Phi(Z) + \sigma(\Xi)\phi(Z) & \text{if } \sigma(\Xi) > 0 \\ 0 & \text{if } \sigma(\Xi) = 0 \end{cases} \qquad \textbf{Eq. (8-5)}$$

where $Z = [R_\chi(\hat{\Xi}) - R_\chi(\Xi)]/\sigma(\Xi)$, $R_\chi(\hat{\Xi})$ is the current minimum value of R_χ among all the known points (in other words, $\hat{\Xi}$ is the current optimal set of parameters), and $\Phi(Z)$ and $\phi(Z)$ are the cumulative distribution and probability density function of the standard normal distribution, respectively. By construction, the value of $EI(\Xi)$ is high (i) when the expected value of $R_\chi(\Xi)$ is smaller than the current best value $R_\chi(\hat{\Xi})$ or (ii) when the uncertainty $\sigma(\Xi)$ around the point $\hat{\Xi}$ is high. Therefore, the maximum position of $EI(\Xi)$ indicates either a point for which a better minimum position of R_χ than the current one is expected or a point belonging to a region of R_χ that has not been explored yet (i.e., $\sigma(\Xi)$ is high). Namely, the maximum position of $EI(\Xi)$ offers the best tradeoff between "exploration" (i.e., minimizing the uncertainty $\sigma(\Xi)$) and "exploitation" (i.e., minimizing the cost function $R_\chi(\Xi)$).

As an illustration of the BO approach, Fig. 8-2(b) shows the computed expected improvement function based on the interpolated function and uncertainty thereof shown in Fig. 8-2(a). As mentioned above, only the partial charge of the Si atoms q_{Si} is here optimized, while the other 9 forcefield parameters are kept fixed and equal to those found in the original BKS potential [12]. As expected, we observe a noticeable maximum in the expected improvement function where the interpolated function R_χ is minimum (exploitation). Some secondary peaks are also observed in the high-uncertainty regions of the function in the vicinity of the minimum position.

(iv) Iterative refinement of the forcefield

Finally, at each step of our iterative optimization scheme, the set of parameters Ξ corresponding to the maximum of the expected improvement function is used to conduct an MD simulation and calculate the associated cost function value R_χ. In turn, this new datapoint is added to the dataset. This enhances the accuracy of the GPR interpolation, which contributes to further refine the sampling of the cost function R_χ at the vicinity of its minimum positions. This iterative

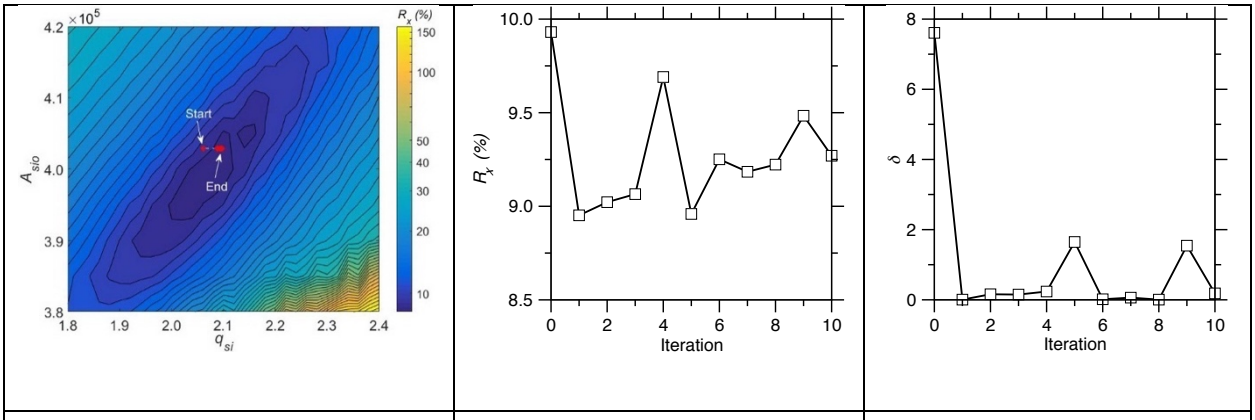
scheme is repeated until convergence is achieved, that is, until the cost function reaches a plateau and does not further decrease within 100 iterations.

This iterative refinement method is illustrated in Fig. 8-3. Here, for illustrative purposes, only two parameters (q_{Si} and A_{SiO}) are optimized, while the other 8 forcefield parameters are kept fixed and equal to those found in the original BKS potential [12]. Figure 8-3(a) shows a contour plot of the cost function R_χ as a function of the two free parameters used in the optimization. We observe that, even in the case of only two free parameters, the cost function shows a rough dependence on the parameters and exhibits two distinct minima (i.e., the dark blue domains in Fig. 8-3(a)). Figure 8-3(a) also shows the pathway that is explored by the optimization algorithm in the (q_{Si}, A_{SiO}) space, that is, the set of parameters for which the expected improvement function is maximum after each step. We observe that the optimization quickly converges toward the global minimum of the cost function after only 5 iterations, after which the cost function R_χ shows a plateau around 10% (see Fig. 8-3(b)). This illustrates the efficiency of our optimization technique.

8.2.4 Final refinement by conjugate gradient (CG)

Finally, the minimum identified by the iterative BO scheme is further refined by the CG method. Indeed, although the BO method can quickly identify the vicinity of the global minimum of a rough function, the CG method is more efficient to pinpoint the minimum position in a local basin of the cost function. Here, we adopt the nonlinear CG algorithm detailed in Ref [16]. In short, we first use the secant method to construct a quadratic interpolation of $R_\chi(\Xi)$ at the vicinity of the minimum identified by the iterative BO scheme and determine the new minimum predicted by the CG interpolation. We then repeat the quadratic construction (i.e., the linear search) around this new minimum position. This is used to approximate the minimum position of R_χ along the CG

direction (i.e., the search direction). The maximum number of iterations of linear search in a search direction is set as 3. Then, starting from the identified new minimum position, we calculate the local gradient and find a new search direction based on Polak-Ribiere formula [16]. A new search direction is then determined from this starting point to identify a new minimum position. The iterative scheme is repeated until convergence, that is, when the new minimum position largely overlaps with the last minimum position, R_χ shows a plateau, and the squared sum of the local gradient converges toward zero and remains lower than the "zero" threshold (taken as 5 herein) within 10 iterations.



(a) **(b)** **(c)**

Figure 8-4: Illustration of the final conjugate gradient optimization. Only the partial charge of the Si atoms q_{Si} and the parameter A_{SiO} are here optimized, while the other 8 forcefield parameters are kept fixed. **(a)** Contour plot showing the cost function R_χ as a function of q_{Si} and A_{SiO}. The white dashed line indicates the path explored by the conjugate gradient optimization method until the minimum in the cost function R_χ is identified. **(b)** Evolution of the cost function R_χ during the conjugate gradient optimization process. **(c)** Evolution of the squared-sum of the local gradient δ during the conjugate gradient optimization process.

Figure 8-4 shows an illustration of the CG refinement step—starting from the minimum identified by the BO iterative scheme illustrated in Fig. 8-3. By exploring "downhill" the local minimum of the cost function (Fig. 8-4(a)), the CG allows us to further refine the position of the minimum—the cost function decreasing from 10% to about 9% (see Fig. 8-4(b)). As expected, the local gradient converges toward zero as the CG optimization proceeds (see Fig. 8-4(c)). Note that, due to the high roughness of the cost function, CG optimization alone cannot yield a satisfactory minimum for the cost function as it easily gets stuck in local minima [26].

8.3 Results

8.3.1 Accuracy of the forcefields

We now assess how the degree of complexity of the forcefield controls its accuracy. To this end, we compare the outcomes of our ML-based parametrization method for three forcefields featuring increasing degrees of complexity (see Sec. 8.2.1), namely, (i) ML-SiO, which only comprises Si–O energy terms, (ii) ML, which comprises Si–O and O–O energy terms (i.e., like the well-established BKS potential [12,27]), and (iii) ML-ALL, which comprises Si–O, O–O, and Si–Si energy terms (i.e., like the CHIK potential [11]). Note that, in all cases, the Coulombic interactions are computed for all the pairs of atoms—so that only the "Buckingham" contribution of these three potentials is varied. In order of increasing complexity, the three potentials comprise 4, 7, and 10 variable parameters, respectively (i.e., 3 parameters per interatomic pair and the Si partial charge). From a physical viewpoint, this analysis allows us (i) to investigate whether accounting for O–O interaction terms (i.e., besides the Coulombic repulsion) is truly necessary to predict a realistic structure for glassy silica and (ii) to assess the extent to which incorporating Si–Si energy terms can improve the performance of the forcefield. More generally, this analysis is

conducted to identify the right level of complexity, that is, to develop a forcefield that is neither underfitted nor overfitted.

The parameters obtained for the ML potential are listed in Tab. 8-1, whereas those obtained for the ML-SiO and ML-ALL potentials are listed in Tab. 8-2 and 8-3. Overall, we find that the parameters of the ML-SiO forcefield significantly differ from those of the ML forcefield. In particular, we obtain a very small Si partial charge of +1.484. In contrast, we note that the parameters of the ML-ALL forcefield are largely similar to those of the ML potential—with a partial charge for Si atoms that is around 1.955. This value is fairly close to that of the CHIK (+1.91 [14]) and Wang–Bauchy potentials (+1.89 [6]).

Figure 8-5 presents a comparison of the accuracy of the three forcefields (as quantified in terms of the final cost function R_χ). We note that the low-complexity ML-SiO potential offers a very poor description of the structure of silica (i.e., high final R_χ value—note that a threshold of 10% is typically used to discriminate "good" from "bad" forcefields [18]). This confirms that, as expected, the O–O interactions play a key role in predicting a realistic SiO_2 structure and that the ML-SiO model is clearly underfitted. In contrast, as shown in Fig. 8-5, the high-complexity ML-ALL potential offers a slight improvement in the description of the structure of silica with respect to that predicted by ML potential, which manifests itself by a slight decrease in R_χ from 8.77% to 7.20%. Although this improvement is higher than the level of uncertainty in the R_χ values, it remains small as compared to the difference between the R_χ values yielded by the ML and ML-SiO forcefields. This suggests that Si–Si interactions only play a minor role in controlling the structure of silica. In turn, this small improvement comes with a significantly higher degree of complexity (i.e., 3 extra parameters), which suggests that the ML-ALL potential may be overfitted.

Table 8-1. Parameters of the optimized potential "ML" (see Eq. (8-1)). The partial charges are indicated as superscripts for each pair of atoms.

Atomic pairs	$Si^{+1.955} – O^{-0.9775}$	$O^{-0.9775} – O^{-0.9775}$	$Si^{+1.955} – Si^{+1.955}$
A (eV)	20453.6 ± 0.2	1003.4 ± 0.2	0
ρ (Å)	0.191735 ± 0.000005	0.356855 ± 0.000005	1
C (eV·Å^6)	93.5 ± 0.5	81.5 ± 0.5	0

Table 8-2. Parameters of the interatomic potential "ML-SiO" (which only considers Si–O interactions). The partial charges are indicated as superscripts for each pair of atoms.

Atomic pairs	$Si^{+1.484} – O^{-0.742}$	$O^{-0.742} – O^{-0.742}$	$Si^{+1.484} – Si^{+1.484}$
A (eV)	3968.5 ± 0.2	0	0
ρ (Å)	0.187600 ± 0.000005	1	1
C (eV·Å^6)	0.7 ± 0.5	0	0

Table 8-3. Parameters of the interatomic potential "ML-ALL" (which includes Si–Si interactions). The partial charges are indicated as superscripts for each pair of atoms.

Atomic pairs	$Si^{+1.955} – O^{-0.9775}$	$O^{-0.9775} – O^{-0.9775}$	$Si^{+1.955} – Si^{+1.955}$
A (eV)	20453.6 ± 0.2	1003.4 ± 0.2	2643.1 ± 0.2
ρ (Å)	0.191735 ± 0.000005	0.356855 ± 0.000005	0.303616 ± 0.000005
C (eV·Å^6)	93.5 ± 0.5	81.5 ± 0.5	232.0 ± 0.5

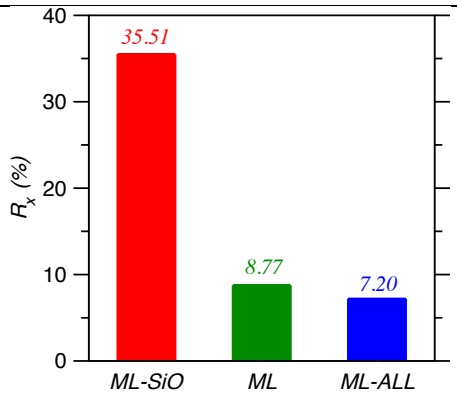
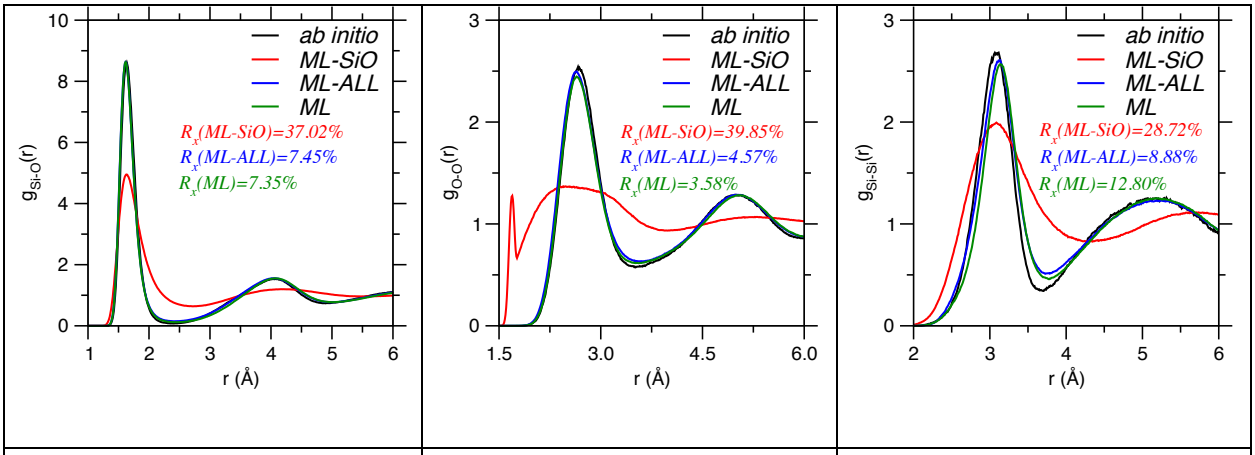


Figure 8-5: Comparison of the final cost function values R_χ obtained by including, in order of increasing complexity: (i) only Si–O interactions ("ML-SiO" potential), (ii) both Si–O and O–O interactions ("ML" potential), and (iii) Si–O, O–O, and Si–Si interactions ("ML-ALL" potential). The relative uncertainty in the R_χ values is ±0.5%, ±0.25%, and ±0.25% for the ML-SiO, ML, and ML-ALL forcefields, respectively.

8.3.2 Partial pair distribution functions

We now further investigate the effect of the complexity of the forcefield on the structure of the simulated liquid silica system (i.e., at 3600 K). To this end, Fig. 8-6 shows a comparison of the partial PDFs obtained by each of the three potentials. The data are compared with the reference *ab initio* partial PDFs used for the training of the potentials. We first focus on the ML potential (i.e., which exhibits the same level of complexity as the BKS potential). Overall, we find that the ML potential provides an excellent agreement with AIMD simulations— although this is not surprising as our forcefield is specifically trained to match these data. Nevertheless, these results illustrate that the Buckingham formulation (see Eq. (8-1)) is adequate to describe the SiO_2 system. This result also further supports the ability of our ML-based optimization method to offer a robust parametrization. We note that the average Si–Si distance predicted by ML potential is slightly

shifted compared with AIMD simulations (see Fig. 8-6(c)). This may arise from a general limitation of the Buckingham formulation.



(a) **(b)** **(c)**

Figure 8-6: (a) Si–O, **(b)** O–O, and **(c)** Si–Si partial pair distribution functions in liquid silica (at 3600 K) predicted by the different forcefields parameterized herein. The data are compared with the *ab initio* reference [14].

We now focus on the low-complexity ML-SiO forcefield (which does not comprise O–O energy terms). Overall, we find that the ML-SiO forcefield exhibits a very unrealistic structure. Although this forcefield succeeds at predicting a reasonable average Si–O average interatomic distance (see Fig. 8-6(a)), it completely fails to properly model O–O correlations (see Fig. 8-6(b)). This confirms that including the O–O interactions is necessary to properly describe the tetrahedral structure of Si atoms and, hence, that the low-complexity ML-SiO forcefield is underfitted.

In contrast, we note that the structure predicted by the ML-ALL forcefield is largely similar to that offered by the ML potential, which confirms that Si–Si interactions play a fairly trivial role in controlling the structure of glassy SiO_2. Although we observe that taking into account Si–Si interactions offers a slight improvement in the Si–Si partial PDF, the average Si–Si distance

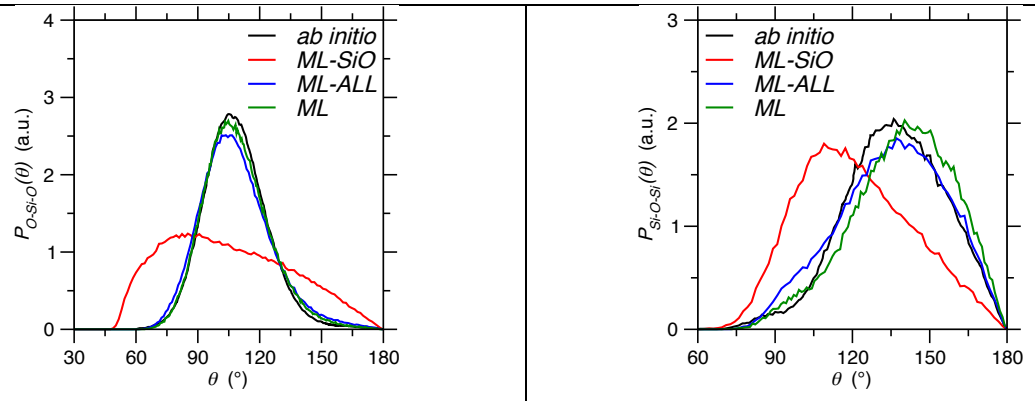
remains overestimated with respect to that predicted by AIMD. This further suggests that this discrepancy is an intrinsic limitation of the two-body Buckingham formulation used herein. Although the inclusion of 3-body energy terms could overcome this limitation, this would come with a significant increase in computing cost and model complexity. Overall, these results confirm that the ML parametrization presented in Tab. 8-1 yields an excellent description of the structure of silica and offers the best balance between accuracy and model simplicity.

8.3.3 Partial bond angle distributions

We now investigate the effect of the forcefield complexity on the partial bond angle distributions (PBADs). Note that, the PBADs are not explicitly included in the cost function (see Eq. (8-2)) and that such 3-body correlations are not fully encoded in the 2-body correlations (i.e., as captured by the partial PDFs). As such, the PBADs allow us to assess the accuracy of the forcefield by comparing their predictions to a structural quantity that is unknown during the training of the forcefields. In that sense, the PBADs acts as a "test set," that is, a group of data that is deliberately kept invisible to the model during parametrization and can then be used to *a posteriori* assess the ability of the model to offer realistic predictions for unknown data. In addition, this analysis allows us to better understand the influence of the O–O and Si–Si interactions in controlling the angular environment of the Si and O atoms.

Figure 8-7 shows the O–Si–O and Si–O–Si PBADs predicted by the three potentials for the liquid silica system (at T = 3600 K). The data are compared with those obtained by *ab initio* simulations [14]. We first focus on the ML potential (i.e., which exhibits the same level of complexity as the BKS potential). Overall, we observe that the PBADs predicted by the ML potential offer a very good agreement with *ab initio* simulations. As expected, the ML potential

yields a tetrahedral environment for Si atoms (with an average O–Si–O angle of about 109°). Nevertheless, we observe that the ML potential slightly overestimates the value of Si–O–Si angles with respect to AIMD simulations, which appears to be a general limitation of the 2-body Buckingham formulation adopted herein and is likely related to the fact that our potential overestimates the Si–Si average distance (see Sec. 8.3.2).



(a) **(b)**

Figure 8-7: (a) O–Si–O and **(b)** Si–O–Si partial bond angle distributions (PBADs) in liquid silica (at $T = 3600$ K) predicted by the different ML-based forcefields parameterized herein. The data are compared with the *ab initio* reference [14].

We now focus on the low-complexity ML-SiO forcefield (which does not comprise O–O energy terms). We note that the ML-SiO potential (i) largely underestimates the average value of both the O–Si–O and Si–O–Si angles and (ii) overestimates the broadness of the angular distributions (i.e., the angular excursions) with respect to the AIMD simulations. These unrealistic PBADs offers a strong *a posteriori* validation of the fact that the ML-SiO potential is underfitted. In turn, these results demonstrate that explicitly accounting for the O–O interactions is essential to correctly model the tetrahedral structure of the Si atoms.

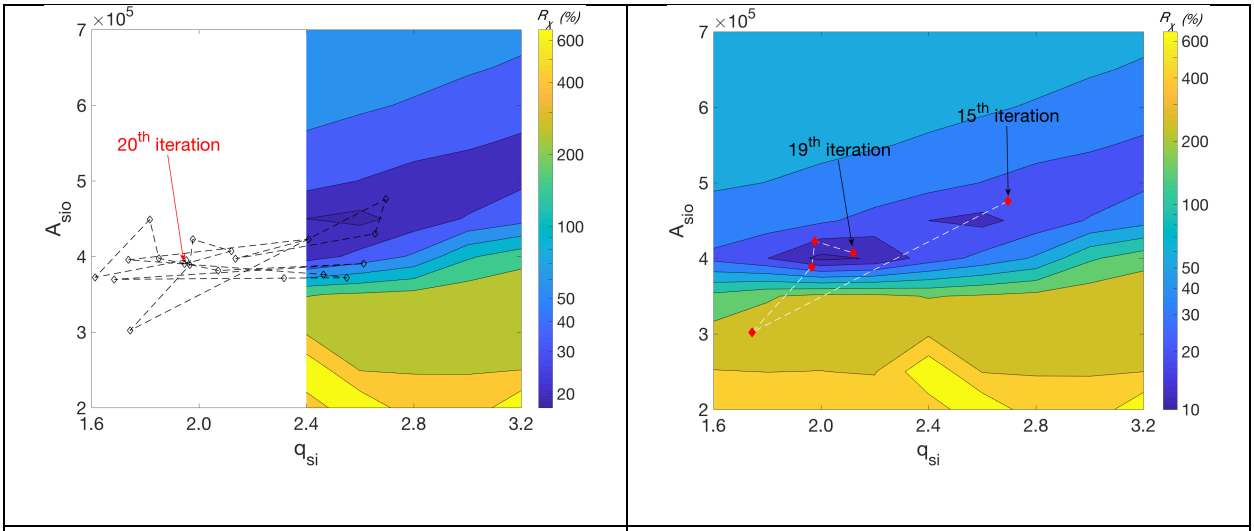
In contrast, we find that the PBADs predicted by the ML-ALL forcefield are fairly similar to those offered by the ML potential, which indicates that accounting for the Si–Si interactions may not be necessary to properly model the angular environment of the Si and O atoms. Further, a more detailed comparison of the PBADs predicted by the ML and ML-ALL potentials with the reference AIMD data reveals that the O–Si–O PBAD predicted by the ML potential is slightly better than that offered by the more complex ML-ALL potential (see Fig. 8-7(a)). Further, we note that, thanks to the addition of Si–Si energy terms, the ML-ALL offers a better description of the average value of the Si–O–Si angle. However, in turn, the Si–O–Si PBAD predicted by the ML-ALL potential exhibit a large degree of asymmetry that is not supported by the AIMD simulations (see Fig. 8-6(c)). This suggests that the fact of capturing all the fine details of the partial PDFs used during the training (as permitted by the high-complexity of the ML-ALL forcefield) results in some overfitting, which, in turn, manifests itself by a decrease in the ability of the potential to properly predict structural metrics that are not explicitly included in the training set. In contrast, due to its higher degree of simplicity, the ML potential only captures the essential features of the partials PDFs and, hence, offers more robust predictions for structural data that are kept invisible during training. This suggests that the ML potential (i.e., which relies on Si–O and O–O energy terms only) presents the best balance between under- and overfitting and, thereby, offers the most accurate overall description of the structure of glassy silica.

8.4 Discussion

8.4.1 Dependence on the initial training set

We now discuss the ability of our ML-based optimization scheme to yield a proper optimal set of forcefield parameters (i.e., to identify a proper minimum in the cost function) regardless of

the choice of the initial training set, that is, the parameter space used as a starting point for the optimization (see Sec. 8.2.3). In particular, it is critical for the parameterization method to be able to yield a minimum in the cost function that is far from the initial training set. Indeed, this is key as we aim to develop a non-biased parameterization scheme that do not rely on "intuition" regarding the range of promising forcefield parameters.



(a) **(b)**

Figure 8-8: Illustration of the iterative optimization approach used herein in the case of the global minimum is far from the initial training set. Only the partial charge of the Si atoms q_{Si} and the parameter A_{SiO} are here optimized, while the other 8 forcefield parameters are kept fixed. In both cases, the contour plot shows the value of the cost function R_χ as a function of q_{Si} and A_{SiO}. The white dashed line indicates the path explored by the Bayesian optimization method until the global minimum in the cost function R_χ is identified. Panel **(a)** highlight in white the parameter space region that is purposely excluded from the initial training set, while panel **(b)** shows the value of cost function over the entire domain to highlight the fact that the global minimum is indeed identified at the end of the optimization.

Figure 8-8 shows an illustration of the ability of our ML-based method to efficiently explore the parameter space—even far away from the initial training set—to yield a proper minimum in the cost function. Here, for illustrative purposes, only two parameters (q_{Si} and A_{SiO}) are optimized, while the other 8 forcefield parameters are kept fixed and equal to those found in the original BKS potential [12]. Figure 8-8(a) shows a contour plot of the cost function R_χ as a function of the two free parameters used in the optimization. In this case, we purposely restrict the region of the initial training set to $q_{\mathrm{Si}} > 2.4$ (i.e., the colored region in Fig. 8-8(a)), which does not comprise the targeted global minimum position of R_χ. We observe that our iterative learning model is able to quickly explore the $q_{\mathrm{Si}} < 2.4$ and identify the global minimum around $q_{\mathrm{Si}} = 2$ despite this position being far from the initial training set (see Fig. 8-8(b)). This signals that the iterative Bayesian optimization is able to "learn" by itself that the global minimum of R_χ does not belong to the initial training set. Overall, these results strongly support the ability of our approach to yield optimal forcefield parameters regardless of the choice of the initial training set considered at the beginning of the parameterization.

8.4.2 Comparison of the ML-based forcefield with previous Buckingham potentials

Finally, we discuss how our new ML potential (i.e., that featuring the optimal degree of complexity) compares with select previous SiO_2 forcefields relying on the Buckingham form. Specifically, we focus on (i) the BKS potential [12], which presents the same complexity as our new ML potential but relies on a different parametrization method and (ii) the CHIK potential [14], which presents a higher complexity (i.e., as it comprises Si–Si energy terms).

Figure 8-9 shows a comparison of the partial PDFs predicted by our new ML forcefield with those predicted by the BKS and CHIK potentials. The data are also compared with the

reference *ab initio* partial PDFs. We observe that both the ML and CHIK potentials offer a clear improvement with respect to the classic BKS potential. Since the ML and BKS forcefield relies on the same formulation and same degree of complexity, these results clearly demonstrate the superiority of our ML-based parametrization method over that used for the BKS potential—which relies on *ab initio* calculations performed on small SiO_4 clusters and the incorporation of some bulk properties during training [12]. On the other hand, we find that our new ML forcefield offers a slightly more accurate prediction of the partial PDFs as compared to the CHIK potential while relying on a lower number of parameters (i.e., lower complexity). This confirms once again that Si–Si interactions are not playing a critical role in governing the structure of glassy SiO_2 and that, in turn, using Si–Si interactions as free parameters during the training of the forcefield can result in some degree of overfitting.

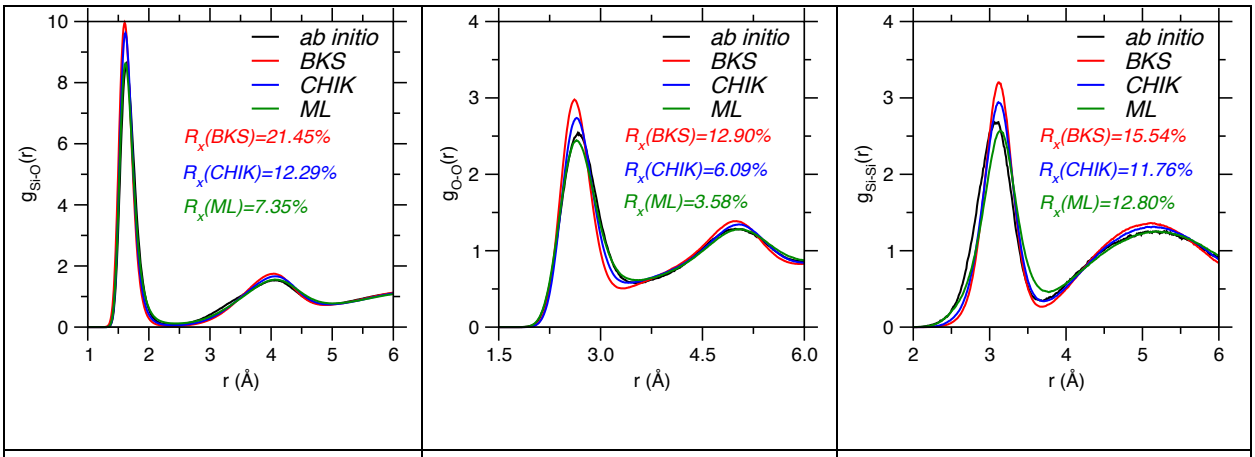


Figure 8-9: (a) Si–O, **(b)** O–O, and **(c)** Si–Si partial pair distribution functions (PDFs) in liquid silica (at $T = 3600$ K) predicted by our new "ML" forcefield and compared with the *ab initio* reference [14]. The partial PDFs predicted by the BKS potential [12] and CHIK potential [14] are added for comparison.

Table 8-4. Unit cell parameters and elastic constants of α-quartz measured by experiments and offered by different Buckingham potentials.

Observable	Experiments [28–30]	BKS	CHIK	ML
V(Å^3)	112.93	119.30 ± 0.06	125.18 ± 0.05	128.8 ± 0.3
a (Å)	4.9124	5.026 ± 0.001	5.1166 ± 0.0008	5.1593 ± 0.0007
c (Å)	5.4039	5.4526 ± 0.0006	5.5212 ± 0.0006	5.5862 ± 0.0005
C_{11} *(GPa)*	86.8	90.9 ± 3.1	98.4 ± 0.6	89.2 ± 5.6
C_{33} *(GPa)*	105.8	116.3 ± 0.6	91.1 ± 0.9	67.0 ± 4.3
C_{44} *(GPa)*	58.2	48.6 ± 0.9	50.3 ± 0.5	46.1 ± 0.4
C_{66} *(GPa)*	39.9	46.0 ± 0.6	43.7 ± 1.1	23.9 ± 4.5
C_{12} *(GPa)*	7.0	–4.0 ± 1.3	–0.2 ± 0.7	–2.8 ± 9.3
C_{13} *(GPa)*	19.1	12.8 ± 0.3	16.8 ± 0.9	8.7 ± 3.6
C_{14} *(GPa)*	–18.0	–0.3 ± 0.1	–0.1 ± 0.2	–0.3 ± 1.0

Finally, we assess whether our new ML forcefield offers a good transferability to α-quartz—that is, whether it can properly describe the structure and stiffness of α-quartz without being explicitly trained for this system. To this end, we compute the unit cell parameters at 300 K and elastic constants at 0 K of α-quartz using our potential (see Tab. 8-4) and compare these values to available experimental data [28–30]. These data are also compared with the values offered by the BKS and CHIK potentials. Overall, we find that our potential reproduces experimental data with a degree of accuracy that is comparable to that offered by previous potentials based on the Buckingham formulation (i.e., BKS and CHIK). This is notable as (i) α-quartz is not part of the training set used for the present ML forcefield and (ii) our forcefield was not explicitly trained to

reproduce any stiffness data. This demonstrates that the pair distribution function (used to train the ML forcefield) contains enough details about the simulated system to offer a realistic description of the curvature of the interatomic potential (which largely controls stiffness). More generally, this shows that our new ML potential shows a satisfactory transferability to new phases (i.e., α-quartz) that are not explicitly considered during training.

8.5 Conclusions

Overall, this study establishes a general and versatile framework to facilitate the development of accurate, yet transferable empirical forcefields for the modeling of disordered materials. By taking the example of silica, our method is able to quickly parameterize forcefields featuring different degrees of complexity in a non-biased fashion. This robust method allows us to meaningfully assess the optimal degree of complexity for the forcefield, that is, for which an optimal balance between accuracy and simplicity is achieved. The assessment of the role of the complexity of forcefields is key to avoid any overfitting, which would likely decrease the transferability of the potential to new systems that are not explicitly included during training. More generally, we expect that the use of ML will decrease the importance of intuition for the parametrization of future potentials for multicomponent silicate glasses.

8.6 References

[1] L. Huang, J. Kieffer, Challenges in Modeling Mixed Ionic-Covalent Glass Formers, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 87–112. doi:10.1007/978-3-319-15675-0_4.

[2] J. Du, Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials, Springer International Publishing, 2015: pp. 157–180.

[3] A.C.T. van Duin, S. Dasgupta, F. Lorant, W.A. Goddard, ReaxFF: A Reactive Force Field for Hydrocarbons, J. Phys. Chem. A. 105 (2001) 9396–9409. doi:10.1021/jp004368u.

[4] Y. Yu, B. Wang, M. Wang, G. Sant, M. Bauchy, Revisiting silica with ReaxFF: Towards improved predictions of glass structure and properties via reactive molecular dynamics, Journal of Non-Crystalline Solids. 443 (2016) 148–154. doi:10.1016/j.jnoncrysol.2016.03.026.

[5] Y. Yu, B. Wang, M. Wang, G. Sant, M. Bauchy, Reactive Molecular Dynamics Simulations of Sodium Silicate Glasses — Toward an Improved Understanding of the Structure, Int J Appl Glass Sci. 8 (2017) 276–284. doi:10.1111/ijag.12248.

[6] M. Wang, N.M. Anoop Krishnan, B. Wang, M.M. Smedskjaer, J.C. Mauro, M. Bauchy, A new transferable interatomic potential for molecular dynamics simulations of borosilicate glasses, Journal of Non-Crystalline Solids. 498 (2018) 294–304. doi:10.1016/j.jnoncrysol.2018.04.063.

[7] L. Deng, J. Du, Development of boron oxide potentials for computer simulations of multicomponent oxide glasses, Journal of the American Ceramic Society. (2018) 1–24. doi:10.1111/jace.16082.

[8] X. Li, W. Song, K. Yang, N.M.A. Krishnan, B. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Balonis, M. Bauchy, Cooling rate effects in sodium silicate glasses: Bridging the gap between molecular dynamics simulations and experiments, The Journal of Chemical Physics. 147 (2017) 074501. doi:10.1063/1.4998611.

[9] J.M.D. Lane, Cooling rate and stress relaxation in silica melts and glasses via microsecond molecular dynamics, Physical Review E. 92 (2015). doi:10.1103/PhysRevE.92.012320.

[10] K. Vollmayr, W. Kob, K. Binder, Cooling-rate effects in amorphous silica: A computer-simulation study, Physical Review B. 54 (1996) 15808–15827. doi:10.1103/PhysRevB.54.15808.

[11] A. Carré, S. Ispas, J. Horbach, W. Kob, Developing empirical potentials from ab initio simulations: The case of amorphous silica, Computational Materials Science. 124 (2016) 323–334. doi:10.1016/j.commatsci.2016.07.041.

[12] B.W.H. van Beest, G.J. Kramer, R.A. van Santen, Force fields for silicas and aluminophosphates based on *ab initio* calculations, Physical Review Letters. 64 (1990) 1955–1958. doi:10.1103/PhysRevLett.64.1955.

[13] P. Ganster, M. Benoit, J.-M. Delaye, W. Kob, Structural and vibrational properties of a calcium aluminosilicate glass: classical force-fields vs. first-principles, Molecular Simulation. 33 (2007) 1093–1103. doi:10.1080/08927020701541006.

[14] A. Carré, J. Horbach, S. Ispas, W. Kob, New fitting scheme to obtain effective potential from Car-Parrinello molecular-dynamics simulations: Application to silica, EPL. 82 (2008) 17001. doi:10.1209/0295-5075/82/17001.

[15] S. Sundararaman, L. Huang, S. Ispas, W. Kob, New optimization scheme to obtain interaction potentials for oxide glasses, J. Chem. Phys. 148 (2018) 194504. doi:10.1063/1.5023707.

[16] J.R. Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, Carnegie Mellon University, 1994.

[17] C.J. Fennell, J.D. Gezelter, Is the Ewald summation still necessary? Pairwise alternatives to the accepted standard for long-range electrostatics, The Journal of Chemical Physics. 124 (2006) 234104. doi:10.1063/1.2206581.

[18] A.C. Wright, The comparison of molecular dynamics simulations with diffraction experiments, Journal of Non-Crystalline Solids. 159 (1993) 264–268. doi:10.1016/0022-3093(93)90232-M.

[19] J.-P. Hansen, I.R. McDonald, Theory of Simple Liquids: with Applications to Soft Matter, Academic Press, 2013.

[20] R. Car, M. Parrinello, Unified Approach for Molecular Dynamics and Density-Functional Theory, Physical Review Letters. 55 (1985) 2471–2474. doi:10.1103/PhysRevLett.55.2471.

[21] N.P. Bansal, R.H. Doremus, Handbook of Glass Properties, Elsevier, 2013.

[22] P. Ganster, M. Benoit, W. Kob, J.-M. Delaye, Structural properties of a calcium aluminosilicate glass from molecular-dynamics simulations: A finite size effects study, J. Chem. Phys. 120 (2004) 10172–10181. doi:10.1063/1.1724815.

[23] P.I. Frazier, J. Wang, Bayesian Optimization for Materials Design, in: Information Science for Materials Discovery and Design, Springer, Cham, 2016: pp. 45–75. doi:10.1007/978-3-319-23871-5_3.

[24] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, 2008.

[25] Y.L. Tong, The Multivariate Normal Distribution, Springer-Verlag, New York, 1990.

[26] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Parameterization of Empirical Forcefields for Glassy Silica using Machine Learning, MRS Communications. (2019).

[27] B. Wang, Y. Yu, Y.J. Lee, M. Bauchy, Intrinsic Nano-Ductility of Glasses: The Critical Role of Composition, Front. Mater. 2 (2015) 11. doi:10.3389/fmats.2015.00011.

[28] G. Will, M. Bellotto, W. Parrish, M. Hart, Crystal structures of quartz and magnesium germanate by profile analysis of synchrotron-radiation high-resolution powder data, Journal of Applied Crystallography. 21 (1988) 182–191. doi:10.1107/S0021889887011567.

[29] L. Levien, C.T. Prewitt, D.J. Weidner, Structure and elastic properties of quartz at pressure, American Mineralogist. 65 (1980) 920–930.

[30] H.J. McSkimin, P. Andreatch, R.N. Thurston, Elastic Moduli of Quartz versus Hydrostatic Pressure at 25° and − 195.8°C, Journal of Applied Physics. 36 (1965) 1624–1632. doi:10.1063/1.1703099.

Chapter 9. Exploring the Landscape of Buckingham Potentials for Silica by Machine Learning: Soft vs Hard Interatomic Forcefields

9.1 Introduction

Molecular dynamics (MD) simulations are now routinely used to access the atomic structure of glasses, which is key to decode the relationship between their composition and properties [1–4]. Starting from an initial configuration, classical MD simulations predict the trajectory of the atoms by numerically solving the Newton's law of motion [1]. Assuming that the timestep is low enough to avoid any spurious effects arising from the numerical integration, the accuracy of classical MD simulation is essentially controlled by that of the empirical interatomic forcefield that is used [5,6].

In general, traditional empirical forcefields are an attempt to simplify the quantum mechanical reality into simpler analytical functions describing interatomic interactions [7,8]—wherein forcefields are optimized in order to maximize the level of agreement between the simulation and given references (e.g., experimental properties or first principle calculations). This process can be rationalized as an optimization process, wherein a given cost function capturing the difference between simulation and references is minimized by adjusting the parameters of these analytical functions [9].

Nevertheless, developing forcefields is a tedious process as such cost functions often feature several, competing minima, that is, several sets of parameters can yield comparable forcefield accuracies [10]. This difficulty can be discussed in terms of the topography of the "landscape of a forcefield," wherein the landscape represents the evolution of the overall forcefield accuracy (i.e., the value of the cost function) as a function of the value of the forcefield parameters. Forcefields landscapes are often non-parabolic and can be multistable, that is, the landscape

features several valleys [11]. Such roughness in forcefield landscapes is a natural consequence of their empirical, partially non-physical nature and can result from the existence of mutually-dependent parameters or counterbalancing effects between parameters. Multistable forcefield landscapes present unique challenges during their parameterization, since the outcome of the parameterization often depends on the choice of the initial parameters (i.e., the initial position in the landscape). This often renders the process of forcefield optimization biased [12]—so that forcefield parameterization is sometimes referred as art rather than science [8,13]. In that regard, special effort has been placed in developing accurate forcefields for glassy silica (SiO_2)—an archetypal model for more complex modified silicate glasses [10,14–19]. It is nevertheless interesting to point out that, despite the apparent simplicity of this system, glassy silica often acts as an outlier and, quite surprisingly, is sometimes more challenging to accurately model as compared to its modified silicate counterparts [6,18,20,21].

Many empirical forcefields have been developed to model silicate glasses—each of them focusing on distinct structure features and properties [14,15,17,10,18,19,22,23]. As ionocovalent systems, silicate phases are typically well described by combining short-range Buckingham-form potentials (see Eq. (9-1) below) with long-range coulombic interactions with fixed charges [16,24–29]. Although early simulations typically adopted formal charges charge (e.g., +4 and –2 for Si and O atoms, respectively), they usually required the use of additional angular 3-body energy terms to properly describe the tetrahedral structure of the SiO_4 units [5,6,30]. However, it is now recognized that the use of 3-body terms can be avoided by relying on partial (rather than formal) charges [5,6,14,15]. Although such partial charges are primarily an additional fitting parameter, they capture underlying charge transfer (due to the partially covalent nature of Si–O bonds) and polarization effects in interatomic bonds [14,15]. Importantly, this development has facilitated

simulations of large-scale systems over extended timescales—since angular 3-body terms are significantly more computationally expensive to compute than radial 2-body terms [31].

In the case of silicate systems, Buckingham forcefields relying on partial charges can be classified as "soft" or "hard" potentials based on the value of partial charge q_{Si} attributed to Si atoms—wherein soft forcefields rely on fairly small partial charges (typically $q_{Si} \approx 2$) whereas hard forcefields use higher charges (typically $q_{Si} \geq 2.4$) [14,16,25–27]. Note that the partial charges of the other atoms are then determined to ensure electronic neutrality, for instance, $q_O = -q_{Si}/2$. Although both soft and strong forcefields have been shown to offer a fairly good representation of silicate glasses [14–16], the exact role played by the value of the partial charges in describing the structure and properties of silicate glasses remains poorly understood. This partially arises from the fact that no systematic exploration of the accuracy of Buckingham forcefields as a function of the partial charge value has been conducted thus far.

Here, for the first time, we systematically explore the landscape of Buckingham forcefields with fixed partial charges for silica. Such a systematic exploration is made possible by benefiting from our recently-developed machine-learning-based forcefield parameterization method, which allows us to efficiently parameterize interatomic forcefields in an unbiased fashion [10,12,32]. We observe that, overall, forcefields relying on partial charges offer an improved accuracy as compared to those based on formal charges. Interestingly, we find that soft and hard forcefields correspond to two distinct, yet competitive local minima in the landscape of Buckingham forcefields for silica. We show that both soft and hard potentials yield an equally accurate description of the short-range order structure of liquid silica. However, we find that soft potentials offer an enhanced description of the medium-range order structure as compared to hard potentials.

9.2 Methods

9.2.1 Buckingham potential

We focus here on the pairwise Buckingham-form empirical potential—a formulation that relies on partial charges and typically provides a good description of ionocovalent systems [14,16,17,22]. The interatomic energy between atoms *i* and *j* is expressed as:

$$U_{ij} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + A_{ij}\exp\left(-\frac{r_{ij}}{\rho_{ij}}\right) - \frac{C_{ij}}{r_{ij}^6} + \frac{D_{ij}}{r_{ij}^{24}} \qquad \textbf{Eq. (9-1)}$$

where r_{ij} is the distance between each pair of atoms, q_i is the partial charge of each atom (q_O and q_{Si} for O and Si atoms, respectively—note that $q_O = -q_{Si}/2$), ε_0 is the dielectric constant, and A_{ij}, ρ_{ij}, C_{ij}, and D_{ij} are some parameters describing the short-range interactions. A cutoff of 8 Å is consistently used for the short-range interactions. The long-range coulombic interactions are calculated by damped shifted force (dsf) model [33] with a damping parameter of 0.25 and a cutoff of 8 Å. The last term of Eq. (9-1) is artificially added to ensure a strong repulsion at short distance, thereby preventing any atomic overlap known as "Buckingham catastrophe" [16]. The value of the D_{ij} parameters are fixed based on Ref. [16] (viz., D_{ij} = 113, 29, and 3423200 eV·Å^{24} for O–O, Si–O, and Si–Si interactions, respectively).

9.2.2 Cost function for forcefield optimization

The parametrization of this potential (Eq. (9-1)) consists in optimizing 10 independent parameters, namely, the partial charge q_{Si} and the short-range parameters $\{A_{ij}, \rho_{ij}, C_{ij}\}$ for each of the three atomic pairs (Si–O, O–O, and Si–Si). This set of parameters is denoted Ξ thereafter. Note that, here, we do not consider any Si–Si interaction energy term (besides their mutual coulombic repulsion), since such terms were found to result in the overfitting of the forcefield [12]. Following Kob and Huang *et al.*, we define the optimization cost function R_χ as the squared difference

between the total pair distribution function (PDF) $g(r)$ obtained by classical MD simulation and a fixed reference *ab initio* MD (AIMD) simulation [15–17]:

$$R_\chi = \sqrt{\frac{\sum_r [g^{\mathrm{MD}}(r) - g^{\mathrm{AIMD}}(r)]^2}{\sum_r [g^{\mathrm{AIMD}}(r)]^2}} \qquad \textbf{Eq. (9-2)}$$

The technical details of the MD and AIMD simulations are provided below.

(i) Classical MD simulation:

Given a set of forcefield parameters Ξ, a classical MD simulation is conducted on the same liquid silica system using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) code [34]. The simulated system comprises 1000 SiO_2 units (3000 atoms) placed in a periodic cubic simulation box of length 35.661 Å—in accordance with the experimental glass density of 2.2 g/cm^3 [35]. We first fully relaxed the configuration for 10 ps at 3600 K in the canonical (*NVT*) ensemble. The total PDF of the simulated systems is then computed based on statistical averaging in a subsequent *NVT* dynamics of 10 ps. A Nosé-Hoover thermostat [36] and a timestep of 1 fs are consistently used for all simulations.

(ii) Reference AIMD simulation:

The "reference" liquid silica structure is prepared by conducting a AIMD run in the *NVT* ensemble with the Nose-Hoover thermostat using the Vienna *Ab initio* Simulation Package (VASP) [36,37]. 67 SiO_2 units (201 atoms) are placed in a periodic cubic simulation box of length 14.4839 Å to match the experimental glass density of 2.2 g/cm^3 [35]. The electronic structure of the atoms is described within the framework of density functional theory. The choice of pseudopotentials for each atom-type, and exchange and correlation functions, are based on the projector augmented wave (PAW) method along with the Perdew-Burke-Ernzerhof (PBE) correlation energy functional [38,39]. A timestep of 0.5 fs and a plane-wave cutoff of 520 eV are used to ensure accurate evaluation of system energy evolution. The initial configuration is a liquid configuration obtained

by classical MD simulation at 3600 K using the well-established van Beest–Kramer–van Santen (BKS) potential [14]. This configuration is then relaxed via AIMD at 3600 K and constant volume for 6 ps [40]—which duration is long enough due to the small relaxation time of the system at such elevated temperature. A subsequent run of 16 ps is then used for statistical averaging, during which we compute relevant structural features.

Both the MD and AIMD simulations serve to compute the total PDF, bond angle distributions, and ring size distribution of the simulated liquid system. Note that, in the case of the ring size distribution, the number of atoms in classical MD is reduced to match that of the AIMD simulation to avoid any spurious size effect. In both cases, the ring size distribution is calculated by enumerating irreducible primitive rings using the RINGS package [41].

9.2.3 Machine learning optimization

We use our recently-developed machine learning (ML) parameterization approach [10,12,32] to identify optimal sets of forcefield parameters Ξ that minimize the cost function $R_\chi\{\Xi\}$. This approach is based on combining Gaussian process regression (GPR) [42] and Bayesian optimization (BO) [43], wherein (i) a GPR model is trained to interpolate the evolution of the cost function R_χ as a function of the forcefield parameters Ξ, (ii) BO is used to "explore and exploit" the function $R_\chi\{\Xi\}$ so as to pinpoint the locations Ξ wherein R_χ is minimum, and (iii) the "true" cost function R_χ associated with the forcefield parameters Ξ predicted by BO are computed by MD and subsequently added to the training set to refine the GPR model. This process is iteratively repeated until convergence. More technical details can be found in Refs. [10] and [12]. Here, we adopt this method to explore the "landscape" of $R_\chi\{\Xi\}$, that is, the series of forcefield parameters Ξ wherein the cost function R_χ exhibits a local (or global) minimum.

9.3 Results and Discussion

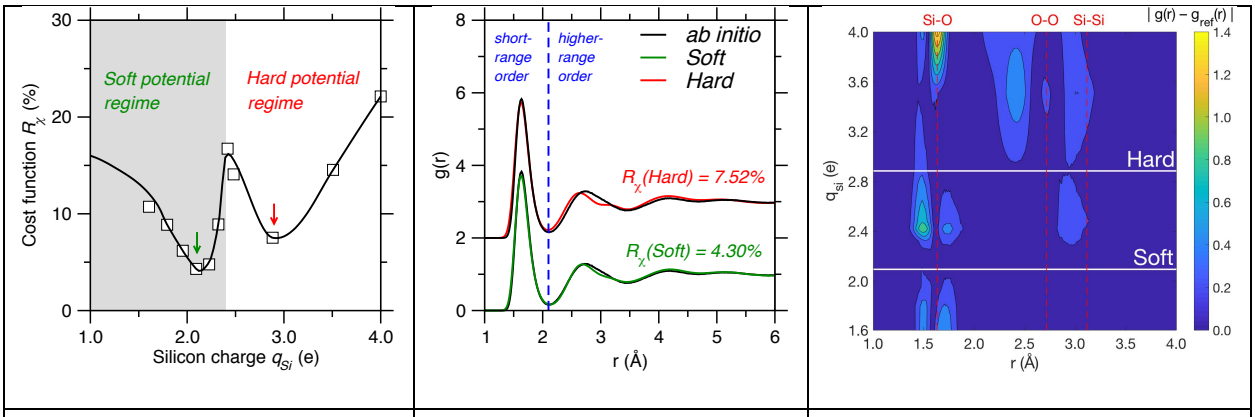
We now explore the topography of the landscape of Buckingham forcefields for silica, that is, how the accuracy of the forcefield (as captured by the cost function R_χ) depends on the value of the forcefield parameters Ξ. As shown in our previous work [10], we emphasize that the landscape of the $R_\chi\{\Xi\}$ is extremely rough, that is, the $R_\chi\{\Xi\}$ function exhibits various local minima that are separated from each other by some barriers. Such roughness renders largely inefficient and biased traditional optimization methods, since such methods (e.g., conjugate gradient) can easily get stuck in local minima—so that the outcome of the minimization depends on the chosen starting point. Each of these local minima corresponds to an optimal set of forcefield parameters Ξ that offers a local maximum in accuracy.

Notably, we find that several sets of forcefield parameters Ξ offer a fairly competitive description of the pair distribution of liquid silica (that is, a competitively small value for R_χ). Figure 9-1(a) shows an unfolded "landscape" of the cost function R_χ as a function of the partial charge attributed to Si atoms q_{Si}. Note that each point in Fig. 9-1(a) corresponds to a local minimum of $R_\chi\{\Xi\}$ identified by our Bayesian optimization approach and that the other parameters of the forcefield $\{A_{ij}, \rho_{ij}, C_{ij}\}$ are different for each point. In particular, we find that several sets of parameters yield to a local minimum featuring $R_\chi < 10\%$, which Is typically considered as a threshold value to discriminate "good" from "bad" forcefields [44]. This high number of competitive local minima is a manifestation of the rough nature of the landscape of Buckingham forcefields for silica and explains why so many forcefields featuring different parameters (and, in particular, different partial charges) have been shown to offer a good description of the structure of silica systems. It is worth pointing out that, overall, the forcefield that is based on formal charges

(i.e., $q_{\mathrm{Si}} = 4$) offers the lowest level of accuracy, which further demonstrates the superiority of forcefields relying on partial charges for silicate systems.

Interestingly, we observe that the landscape exhibits two main basins, which are centered around $q_{Si} \approx 2$ and 3, respectively. These results offer for the first time a quantitative foundation behind the soft vs. hard classification of Buckingham forcefields for silicates. In the following, we refer to "soft" and "hard" forcefields those that are characterized by $q_{\mathrm{Si}} < 2.4$ (i.e., weak coulombic interactions) and $q_{\mathrm{Si}} > 2.4$ (i.e., strong coulombic interactions), respectively. Specifically, we focus in the following on the two soft and hard forcefield parameterizations offering maximum accuracy, that is, for $q_{\mathrm{Si}} = 2.094$ ("soft potential," see Tab. 9-1) and $q_{Si} = 2.883$ ("hard potential," see Tab. 9-2) to further investigate the nature of the structural signatures (if any) behind the soft vs. hard classification.

Figure 9-1(b) shows the total PDFs of silica liquids generated by the soft and hard forcefields. Each PDF is compared to the same *ab initio* reference PDF to assess their accuracy. Overall, we find that both families of forcefield yield a PDF that is an excellent agreement with the AIMD reference PDF, that is, featuring $R_\chi = 4.30\%$ (soft) and 7.52% (hard), respectively. Nevertheless, in both cases, we observed some slight discrepancies. To finely characterize the level of accuracy offered by soft and hard forcefields, Fig. 9-1(c) shows as a contour plot of the absolute error between (i) the PDF generated by classical MD $g(r)$ (for each of the parameterizations presented in Fig. 9-1(a)) and (ii) the same AIMD reference PDF $g_{\mathrm{ref}}(r)$ as a function of the correlation distance r (x-axis) and Si partial charge (y-axis). We observe the existence of "islands" wherein the inaccuracy of the forcefield is locally maximum, which signals that the error associated to different forcefields are mostly concentrated around select correlation distances.



(a) **(b)** **(c)**

Figure 9-1: (a) Cost function R_χ yielded by select Buckingham forcefields for liquid silica as a function of the partial charge of Si atoms q_{Si}. The black line is to guide the eye. The grey and white windows ($q_{Si} < 2.4$ and $q_{Si} > 2.4$) herein defines the soft and hard forcefield regimes, respectively. The arrows indicate the minimum position of R_χ in the soft and hard forcefield regimes (i.e., $q_{Si} = 2.094$ and 2.883, respectively), defining the soft and hard potentials used the following. **(b)** Total pair distribution function (PDF) $g(r)$ of liquid silica at 3600 K generated by the soft and hard potentials offering minimum R_χ. These two PDFs are compared to that generated by *ab initio* molecular dynamics. The blue dashed line ($r = 2.1$ Å) indicates the boundary between the first and second coordination shells, which is in the following used as a threshold distance to define the "short-range order" ($r < 2.1$ Å) and the "higher-range order" ($r > 2.1$ Å). **(c)** Contour plot showing the absolute error between the total PDF $g(r)$ generated by the classical forcefields presented in panel (a) and the *ab initio* reference PDF $g_{ref}(r)$ as a function of the correlation distance r (x-axis) and Si partial charge (y-axis). The horizonal white lines indicate the position of the soft ($q_{Si} = 2.094$) and hard potentials ($q_{Si} = 2.883$). The vertical red lines indicate the values of the average Si–O ($r = 1.635$ Å), O–O ($r = 2.715$ Å), and Si–Si (r = 3.115 Å) interatomic bond distances.

Table 9-1. Parameters of the soft potential. Partial charges are indicated as superscripts for each atom.

Atomic pairs	$Si^{+2.094} – O^{-1.047}$	$O^{-1.047} – O^{-1.047}$	$Si^{+2.094} – Si^{+2.094}$
A (eV)	17471.7 ± 0.2	1386.9 ± 0.2	0
ρ (Å)	0.205205 ± 0.000005	0.362319 ± 0.000005	1
C (eV·Å^6)	133.4 ± 0.5	174.8 ± 0.5	0

Table 9-2. Parameters of the hard potential. Partial charges are indicated as superscripts for each atom.

Atomic pairs	$Si^{+2.883} – O^{-1.4415}$	$O^{-1.4415} – O^{-1.4415}$	$Si^{+2.883} – Si^{+2.883}$
A (eV)	5353.6 ± 0.2	1245.4 ± 0.2	0
ρ (Å)	0.237123 ± 0.000005	0.322535 ± 0.000005	1
C (eV·Å^6)	62.2 ± 0.5	60.0 ± 0.5	0

We first note that the forcefield based on formal charges (i.e., $q_{Si} = 4$) crosses several error islands. In particular, we observe that this potential yields a significant error around $r = 1.6$ Å, that is, around the average Si–O bond distance. This signals that the forcefield based on formal charges is intrinsically unable to offer a good description of the short-range structural order around Si atoms. This confirms that Buckingham potentials relying on formal charges require the use of 3-body angular terms to properly describe the structure of SiO_4 tetrahedra [5,14,15,30].

At lower partial charges ($q_{Si} < 4$), we observe that most of the error islands are located in between the coordination shells, that is, at distances that are slightly lower or higher than the average Si–O, O–O, and Si–Si bond distances (see vertical red dashed lines in Fig. 9-1(c)). This

indicates that, although forcefields relying on partial charges are in general able to properly predict the average positions of the peaks in the total PDF, some level of inaccuracy is observed on each edge of these peaks. This signals that, although the average bond length is a structural feature that is easily reproduced by all forcefields, the shape of each PDF peak (broadness, degree of asymmetry, etc.) is a more sensitive structural feature that can be used to discriminate "good" from "less good" forcefields. It is also noticeable that most of the error islands are located at short distance (around the average Si–O bond length) in the soft regime ($q_{Si} < 2.4$), whereas additional error islands are found at larger distance (around the average O–O and Si–Si bond lengths) in the hard regime ($q_{Si} > 2.4$). Importantly, we find that both the soft and strong forcefields offering maximum accuracy (indicated by white lines in Fig. 9-1(c)) largely avoid any error island, which explains why these two potentials correspond to two distinct local minima in the landscape of Buckingham forcefields for silica.

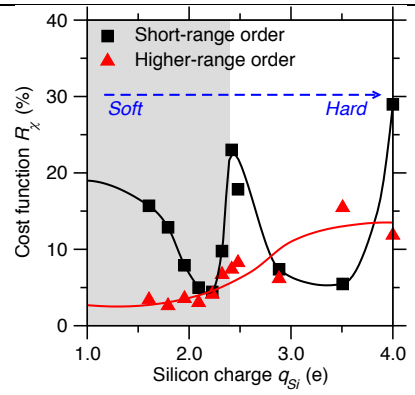
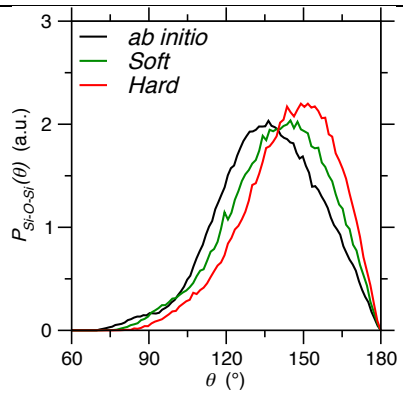
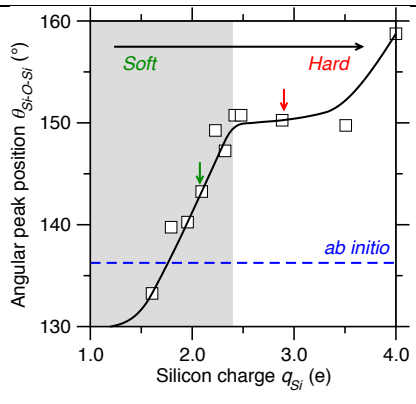


Figure 9-2: Cost function R_χ calculated over the short-range ($r < 2.1$ Å) and higher-range orders ($r > 2.1$ Å) yielded by select Buckingham forcefields for liquid silica as a function of the partial charge of Si atoms q_{Si}. The solid line is to guide the eye. The grey and white windows ($q_{Si} < 2.4$ and $q_{Si} > 2.4$) represent the soft and hard forcefield regimes, respectively.

We now further compare the ability of soft and hard potentials to properly describe the structure of silica at different scales. Based on the location of the error islands in Fig. 9-1(c), we define the "short-range order" (r < 2.1 Å) and the "higher-range order" (r > 2.1 Å) as being the ranges of correlation distances that are lower or higher than the boundary between the first and second coordination shells, respectively (see Fig. 9-1(b)). Figure 9-2 shows the partial cost function R_χ calculated over the short-range and higher-range correlation distance domains (rather than the entire distance domain in Fig. 9-1(a)) as a function of the Si partial charge q_{Si}. Interestingly, we find that the short- and higher-range R_χ metrics exhibit a significantly different trend as a function of q_{Si}. First, we observe that the short-range R_χ metric shows two distinct minima in the soft and hard domains. In this range of distances, the soft and hard forcefields offer a competitive accuracy. This result suggests that both soft and hard potentials relaying on partial charges are able to yield an accurate description at short-range order within the SiO_4 units. In contrast, we observe that the higher-range R_χ metric monotonically increases with increasing q_{Si}, which indicates that, in general, soft potentials offer a more accurate description of higher-distance structural correlation as compared to hard potentials. This finding echoes the fact that the recently proposed Carré–Horbach–Ispas–Kob (CHIK) potential (i.e., a soft Buckingham potential with q_{Si} = 1.91 [15,16]) offers an improved description of liquid silica as compared to the traditional BKS potential (i.e., a fairly hard Buckingham potential with q_{Si} = 2.4 [14]). Overall, these results show that the structural order at distance larger than the first coordination shell is more sensitive to the quality of the forcefield than the short-range structural order. This also suggests that forcefield parameterization methods relying only on short-range order information (e.g., isolated clusters) may not be appropriate to properly describe the structure of silicate glasses.





(a) **(b)**

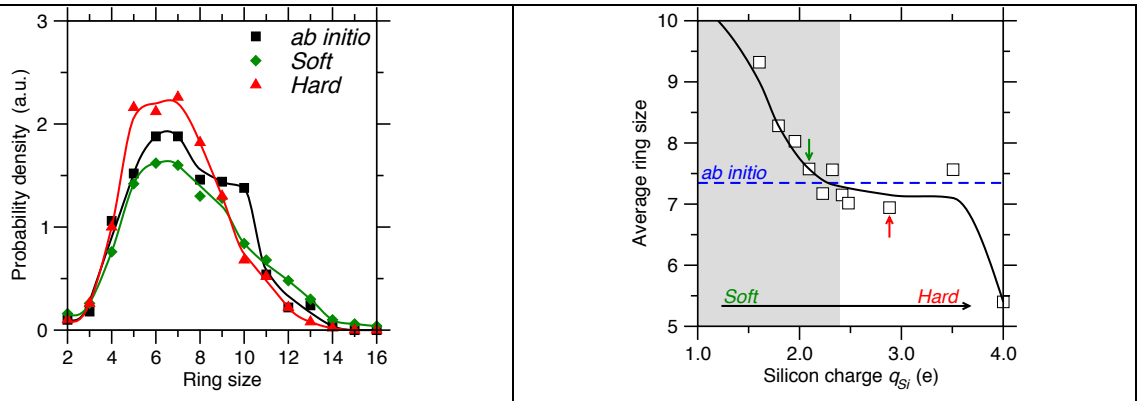
Figure 9-3: (a) Si–O–Si partial bond angle distributions (PBADs) $P_{Si–O–Si}(\theta)$ generated by the soft (q_{Si} = 2.094) and hard potentials (q_{Si} = 2.883). The data are compared with the reference PBAD yielded by *ab initio* molecular dynamics. **(b)** Si–O–Si angular peak position $\theta_{Si–O–Si}$ as a function of the Si partial charge q_{Si}. The black line is to guide the eye. The blue dashed line ($\theta_{Si–O–Si}$ = 136.25°) indicate the *ab initio* reference value. The grey and white windows (q_{Si} < 2.4 and q_{Si} > 2.4) represent the soft and hard forcefield regimes, respectively.

The fact that soft and hard forcefields yield a fairly similar description of Si–O correlations but differ in their description of O–O and Si–Si correlations suggests that these potentials offer different bond angle distributions. We first find that the intra-tetrahedral O–Si–O angle (not shown here) remains fairly unaffected by the value of the Si partial charge. In turn, the inter-tetrahedral Si–O–Si angle presents a higher sensitivity on the forcefield parameterization. As shown in Fig. 9-3(a), we find that both the soft and hard potentials tend to overestimate the average Si–O–Si angle. However, we observe that the soft potential nevertheless offers an improved description of the Si–O–Si partial bond angle distribution as compared to the hard potential. In detail, Figure 9-3(b) shows the Si–O–Si angular peak position $\theta_{Si–O–Si}$ as a function of the Si partial charge q_{Si}. We observe that $\theta_{Si–O–Si}$ increases monotonically with increasing q_{Si}. This can be understood from the

fact that harder forcefields tend to favor the opening of the Si–O–Si angle due to the more intense Si–Si columbic repulsion. Overall, we find that, when compared with AIMD data ($\theta_{\text{Si–O–Si}}$ = 136.25°), hard potentials tend to overestimate the degree of opening of the Si–O–Si angle—which explains the improved ability of soft potentials to describe high-distance structural correlations in silica (see Fig. 9-2).

Finally, we investigate how the Si partial charge affects the medium-range order structure of silica, which is captured by the ring size distribution. Figure 9-4(a) shows the ring size distribution generated by the soft and hard potentials, wherein the size of each ring represents the number of Si the ring is made of. Both distributions are compared with that obtained by AIMD. Overall, we find that both the soft and hard forcefields offer a reasonable description of the ring size distribution. However, we note that the ring distribution generated by the hard potential is sharper and centered around lower ring size as compared to the reference AIMD data. In detail, Fig. 9-4(b) shows the average ring size predicted by each forcefield as a function of the Si partial charge q_{Si}. We observe that the average ring size decreases upon increasing q_{Si}—that is, a trend that is opposite to that observed for the average Si–O–Si angle (see Fig. 9-3). This can be understood as follows. At constant Si–O bond length and constant ring size (in terms of number of Si atoms), the increase in the average Si–O–Si angle upon increasing q_{Si} would result in an increase in the average ring diameter (since the average distance in between pairs of Si atoms increases). Since the density is here fixed, this increase in the average ring diameter upon increasing q_{Si} at fixed ring size must be compensated by a decrease in ring size—so that the volumic density of Si atoms remains constant. Overall, we find that, when compared with AIMD data (average ring size = 7.35), very hard potentials tend to underestimate the average ring size, whereas, in contrast, very soft potentials tend to overestimate the average ring size. Altogether,

moderately soft potentials ($q_{Si} \approx 2$) offer the best description of the reference ring size distribution predicted by AIMD—which suggests that moderately soft potentials exhibit an optimal ability to properly describe the medium-range order structure of silica.



(a) **(b)**

Figure 9-4: (a) Ring size distribution generated by the soft ($q_{Si} = 2.094$) and hard potentials ($q_{Si} = 2.883$). The data are compared with the reference data yielded by *ab initio* molecular dynamics. The lines are to guide the eye. **(b)** Average ring size as a function of the Si partial charge q_{Si}. The solid line is to guide the eye. The blue dashed line (average ring size of 7.35) indicates the reference value yielded by *ab initio* molecular dynamics. The grey and white windows ($q_{Si} < 2.4$ and $q_{Si} > 2.4$) represent the soft and hard forcefield regimes, respectively.

We further assess the ability of the hard and soft potentials to properly describe the medium-range order structure of SiO_2 by computing the neutron structure factor [45]. Indeed, although the structure factor contains the same information as the pair distribution function, it places more emphasis on the medium-range order. Figure 9-5 shows the neutron structure factor $S(Q)$ offered by the soft and hard potentials. Both data are compared with the reference neutron structure factor yielded by *ab initio* molecular dynamics. We observe that, in the short-range order (i.e., high-Q range, $Q > 3$ Å^{-1}), both the soft and hard potentials present a very good agreement

with the *ab initio* reference. However, we find that the soft potential features a notably improved description of the medium-range order (i.e., low-Q range, $Q < 3$ Å^{-1}) as compared to the hard potential. Overall, these results further demonstrate that, although both the soft and hard potentials present a competitive ability to describe the short-range order structure of silica, the soft potential offers a more realistic description of its medium-range order structure.

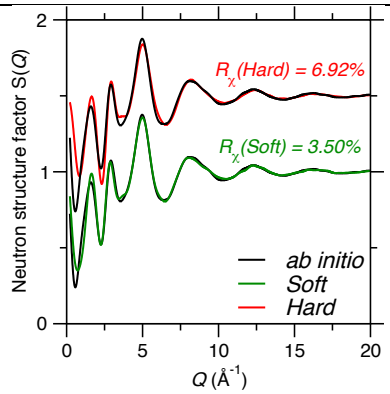


Figure 9-5: Neutron structure factors $S(Q)$ yielded by the soft ($q_{Si} = 2.094$) and hard potentials ($q_{Si} = 2.883$). The data are compared with the reference $S(Q)$ obtained from *ab initio* molecular dynamics.

Finally, we discuss the potential origin of the "bistability" observed herein in the forcefield landscape, namely, the fact that the landscape features two distinct, yet fairly competitive minima (see Fig. 9-1(a)). First, we note that, in the present case, the parameters of the Buckingham potential are not mutually dependent. The two optimal parameterizations identified herein (i.e., hard and soft) yield two distinct overall potential shapes—that is, the landscape bistability cannot be understood in terms of counterbalancing effect between parameters. This is a consequence of the fact that, due to their significantly different shapes, variations in the short-range Buckingham interactions (i.e., exponential and van der Waals terms) cannot be fully counterbalanced by some

variations in the Coulombic term. Rather, we hypothesize that the present bistability arises from the simple 2-body formulation of the Buckingham potential.

Indeed, an accurate forcefield for SiO_2 must yield accurate values for the O–Si–O and Si–O–Si angles. In the case of a 2-body potential like the present one, these angles are not directly constrained but, rather, their value is fixed via a combination of 2-body interactions. Although the average value of the O–Si–O intratetrahedral angle (i.e., 109°) is easily fixed from the strong mutual Coulombic repulsion between the O neighbors, the average value of the intertetrahedral Si–O–Si angle is more sensitive to the quality of the potential. The average value of the Si–O–Si angle is governed by a subtle competition between Si–Si Coulombic repulsion and Si–O long-range attraction (i.e., in between neighboring SiO_4 tetrahedra). Based on this competition, an accurate forcefield must exhibit an optimal balance between its abilities (i) to properly describe the short-range structure of SiO_4 tetrahedra and (ii) to offer a realistic description of longer-range intertetrahedra interactions. The fact that the Buckingham potential relies on a fixed analytical form does not make it possible to independently tune short- and longer-range interactions, which may explain the fact that several sets of parameters will offer competitive accuracies—that is, by achieving different balances between the abilities to properly describe the short- and medium-range structure of SiO_2. Based on this idea, we hypothesize that the "degeneracy" in the forcefield parameterization could be overcome by inserting some additional 3-body terms, which would offer an additional degree of freedom to independently tune intra- and inter-SiO_4 tetrahedra interactions. Nevertheless, the insertion of additional forcefield terms comes with a cost, since it necessarily increases the complexity of the forcefield and associated computing burden, and can also result in a risk of overfitting—which, in turn, can impact the transferability of the forcefield [12]. Additional studies are clearly needed to rigorously explore these hypotheses.

9.4 Conclusions

Overall, these results further confirm that, in the absence of any 3-body angular energy terms, empirical forcefields relying on partial charges offer an improved description of the atomic structure of silica as compared to forcefields relying on formal charges—so that the use of partial charges offers a critical degree of freedom to parameterize accurate interatomic forcefields for silicate systems. To this end, our machine-learning-based parameterization approach offers an efficient route to systematically sample the landscape of Buckingham potentials with fixed partial charges for silica. As a major outcome of this work, we find that such potentials can be formally divided into soft and hard based on the value of the partial charges—wherein these two families of forcefield occupy two distinct basins in the landscape of Buckingham potentials for silica. Although, when properly parameterized, soft and hard potentials offer a competitive description of the short-range order structure, we argue that soft potentials feature an enhanced ability to realistically describe the medium-range order of silica.

9.5 References

[1] C. Massobrio, ed., Molecular dynamics simulations of disordered materials: from network glasses to phase-change memory allyos, Springer, Cham Heidelberg, 2015.

[2] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: A review, Journal of Non-Crystalline Solids: X. 4 (2019) 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[3] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[4] M. Bauchy, Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: A review, Computational Materials Science. 159 (2019) 95–102. https://doi.org/10.1016/j.commatsci.2018.12.004.

[5] J. Du, Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 157–180. https://doi.org/10.1007/978-3-319-15675-0_7.

[6] L. Huang, J. Kieffer, Challenges in Modeling Mixed Ionic-Covalent Glass Formers, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 87–112. https://doi.org/10.1007/978-3-319-15675-0_4.

[7] J.A. Harrison, J.D. Schall, S. Maskey, P.T. Mikulski, M.T. Knippenberg, B.H. Morrow, Review of force fields and intermolecular potentials used in atomistic computational materials research, Applied Physics Reviews. 5 (2018) 031104. https://doi.org/10.1063/1.5020808.

[8] D.W. Brenner, The Art and Science of an Analytic Potential, Physica Status Solidi (b). 217 (2000) 23–40. https://doi.org/10.1002/(SICI)1521-3951(200001)217:1<23::AID-PSSB23>3.0.CO;2-N.

[9] P. Comba, R. Remenyi, Inorganic and bioinorganic molecular mechanics modeling—the problem of the force field parameterization, Coordination Chemistry Reviews. 238–239 (2003) 9–20. https://doi.org/10.1016/S0010-8545(02)00286-2.

[10] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Parameterization of empirical forcefields for glassy silica using machine learning, MRS Communications. (2019) 1–7. https://doi.org/10.1557/mrc.2019.47.

[11] E. Iype, M. Hütter, A.P.J. Jansen, S.V. Nedea, C.C.M. Rindt, Parameterization of a reactive force field using a Monte Carlo algorithm, Journal of Computational Chemistry. 34 (2013) 1143–1154. https://doi.org/10.1002/jcc.23246.

[12] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Balance between accuracy and simplicity in empirical forcefields for glass modeling: Insights from machine learning, Journal of Non-Crystalline Solids. 515 (2019) 133–142. https://doi.org/10.1016/j.jnoncrysol.2019.04.020.

[13] J. Wang, P.A. Kollman, Automatic parameterization of force field by systematic search and genetic algorithms, Journal of Computational Chemistry. 22 (2001) 1219–1228. https://doi.org/10.1002/jcc.1079.

[14] B.W.H. van Beest, G.J. Kramer, R.A. van Santen, Force fields for silicas and aluminophosphates based on *ab initio* calculations, Physical Review Letters. 64 (1990) 1955–1958. https://doi.org/10.1103/PhysRevLett.64.1955.

[15] A. Carré, J. Horbach, S. Ispas, W. Kob, New fitting scheme to obtain effective potential from Car-Parrinello molecular-dynamics simulations: Application to silica, EPL. 82 (2008) 17001. https://doi.org/10.1209/0295-5075/82/17001.

[16] A. Carré, S. Ispas, J. Horbach, W. Kob, Developing empirical potentials from ab initio simulations: The case of amorphous silica, Computational Materials Science. 124 (2016) 323–334. https://doi.org/10.1016/j.commatsci.2016.07.041.

[17] S. Sundararaman, L. Huang, S. Ispas, W. Kob, New optimization scheme to obtain interaction potentials for oxide glasses, J. Chem. Phys. 148 (2018) 194504. https://doi.org/10.1063/1.5023707.

[18] S. Izvekov, B.M. Rice, A new parameter-free soft-core potential for silica and its application to simulation of silica anomalies, The Journal of Chemical Physics. 143 (2015) 244506. https://doi.org/10.1063/1.4937394.

[19] Y. Yu, B. Wang, M. Wang, G. Sant, M. Bauchy, Revisiting silica with ReaxFF: Towards improved predictions of glass structure and properties via reactive molecular dynamics, Journal of Non-Crystalline Solids. 443 (2016) 148–154. https://doi.org/10.1016/j.jnoncrysol.2016.03.026.

[20] L. Huang, J. Kieffer, Amorphous-amorphous transitions in silica glass. I. Reversible transitions and thermomechanical anomalies, Phys. Rev. B. 69 (2004) 224203. https://doi.org/10.1103/PhysRevB.69.224203.

[21] M.S. Shell, P.G. Debenedetti, A.Z. Panagiotopoulos, Molecular structural order and anomalies in liquid silica, Phys. Rev. E. 66 (2002) 011202. https://doi.org/10.1103/PhysRevE.66.011202.

[22] M. Wang, N.M.A. Krishnan, B. Wang, M.M. Smedskjaer, J.C. Mauro, M. Bauchy, A new transferable interatomic potential for molecular dynamics simulations of borosilicate glasses, Journal of Non-Crystalline Solids. 498 (2018) 294–304. https://doi.org/10.1016/j.jnoncrysol.2018.04.063.

[23] A. Pedone, G. Malavasi, M.C. Menziani, A.N. Cormack, U. Segre, A New Self-Consistent Empirical Interatomic Potential Model for Oxides, Silicates, and Silica-Based Glasses, J. Phys. Chem. B. 110 (2006) 11780–11795. https://doi.org/10.1021/jp0611018.

[24] L. Deng, J. Du, Development of boron oxide potentials for computer simulations of multicomponent oxide glasses, Journal of the American Ceramic Society. 102 (2019) 2482–2505. https://doi.org/10.1111/jace.16082.

[25] O. Gedeon, Molecular dynamics of vitreous silica — Variations in potentials and simulation regimes, Journal of Non-Crystalline Solids. 426 (2015) 103–109. https://doi.org/10.1016/j.jnoncrysol.2015.07.006.

[26] A. Takada, P. Richet, C.R.A. Catlow, G.D. Price, Molecular dynamics simulations of vitreous silica structures, Journal of Non-Crystalline Solids. 345–346 (2004) 224–229. https://doi.org/10.1016/j.jnoncrysol.2004.08.247.

[27] T.F. Soules, G.H. Gilmer, M.J. Matthews, J.S. Stolken, M.D. Feit, Silica molecular dynamic force fields—A practical assessment, Journal of Non-Crystalline Solids. 357 (2011) 1564–1573. https://doi.org/10.1016/j.jnoncrysol.2011.01.009.

[28] B.J. Cowen, M.S. El-Genk, On force fields for molecular dynamics simulations of crystalline silica, Computational Materials Science. 107 (2015) 88–101. https://doi.org/10.1016/j.commatsci.2015.05.018.

[29] J. Du, A.N. Cormack, The medium range structure of sodium silicate glasses: a molecular dynamics simulation, Journal of Non-Crystalline Solids. 349 (2004) 66–79. https://doi.org/10.1016/j.jnoncrysol.2004.08.264.

[30] L.V. Woodcock, C.A. Angell, P. Cheeseman, Molecular dynamics studies of the vitreous state: Simple ionic systems and silica, The Journal of Chemical Physics. 65 (1976) 1565–1577. https://doi.org/10.1063/1.433213.

[31] A. Yaseen, H. Ji, Y. Li, A load-balancing workload distribution scheme for three-body interaction computation on Graphics Processing Units (GPU), Journal of Parallel and Distributed Computing. 87 (2016) 91–101. https://doi.org/10.1016/j.jpdc.2015.10.003.

[32] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Machine Learning Forcefield for Silicate Glasses, ArXiv:1902.03486 [Cond-Mat]. (2019). http://arxiv.org/abs/1902.03486 (accessed March 3, 2019).

[33] C.J. Fennell, J.D. Gezelter, Is the Ewald summation still necessary? Pairwise alternatives to the accepted standard for long-range electrostatics, The Journal of Chemical Physics. 124 (2006) 234104. https://doi.org/10.1063/1.2206581.

[34] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[35] N.P. Bansal, R.H. Doremus, Handbook of Glass Properties, Elsevier, 2013.

[36] S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, J. Chem. Phys. 81 (1984) 511–519. https://doi.org/10.1063/1.447334.

[37] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B. 54 (1996) 11169–11186. https://doi.org/10.1103/PhysRevB.54.11169.

[38] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B. 59 (1999) 1758–1775. https://doi.org/10.1103/PhysRevB.59.1758.

[39] D. Hobbs, G. Kresse, J. Hafner, Fully unconstrained noncollinear magnetism within the projector augmented-wave method, Phys. Rev. B. 62 (2000) 11556–11570. https://doi.org/10.1103/PhysRevB.62.11556.

[40] R.M. Van Ginhoven, H. Jónsson, L.R. Corrales, Silica glass structure generation for *ab initio* calculations using small samples of amorphous silica, Phys. Rev. B. 71 (2005) 024208. https://doi.org/10.1103/PhysRevB.71.024208.

[41] S. Le Roux, P. Jund, Ring statistics analysis of topological networks: New approach and application to amorphous GeS2 and SiO2 systems, Computational Materials Science. 49 (2010) 70–83. https://doi.org/10.1016/j.commatsci.2010.04.023.

[42] C.E. Rasmussen, C.K.I. Williams, Gaussian processes for machine learning, 3. print, MIT Press, Cambridge, Mass., 2008.

[43] P.I. Frazier, J. Wang, Bayesian Optimization for Materials Design, in: Information Science for Materials Discovery and Design, Springer, Cham, 2016: pp. 45–75. https://doi.org/10.1007/978-3-319-23871-5_3.

[44] A.C. Wright, The comparison of molecular dynamics simulations with diffraction experiments, Journal of Non-Crystalline Solids. 159 (1993) 264–268. https://doi.org/10.1016/0022-3093(93)90232-M.

[45] M. Bauchy, Structural, vibrational, and elastic properties of a calcium aluminosilicate glass from molecular dynamics simulations: The role of the potential, The Journal of Chemical Physics. 141 (2014) 024507. https://doi.org/10.1063/1.4886421.

Section C. Integration of Machine Learning and Simulations: Toward Next-Generation Materials Modeling

C3. Gain New Physics Knowledge: Deciphering Complex Simulation Data by Machine Learning

Chapter 10. Finding Needles in Haystacks: Deciphering a Structural Signature of Glass Dynamics by Machine Learning

10.1 Introduction

The origin and nature of glass dynamics—i.e., the dynamic motion of the atoms in the glassy state—have remained mysterious for centuries [1–3]. A prominent example of this mystery is manifested as the ubiquitous-yet-indefinite relaxation behaviors of glasses at room temperature [4–6]. Indeed, the dynamics of the atoms governs various dynamical and transport properties of the glass [7,8], including viscosity [9,10], thermal conductivity [11,12], ion diffusivity [13,14], etc. In that regard, understanding the key structural features that control atom dynamics would facilitate the rational design of "tailored" glasses [15,16]. However, due to the complex and disordered nature of glass structures [17,18], pinpointing which structural features (if any) govern dynamics is essentially a "needle-in-a-haystack" problem [19–21], since intuitive structural metrics (e.g., local packing or coordination number) are often only weakly correlated with dynamics [22–24]. As a result, a long-standing debate exists in whether glass dynamics is in some way encoded in the static glass structure[25].

As an emergent thrust to discover hidden patterns in complex, multidimensional data [26–28], machine learning (ML) has become a new paradigm to unveil the nature of the linkages between glass dynamics and its static structure—without the need for any prerequisite intuition regarding which structural feature(s) could be influential [25,29–31]. In particular, Cubuk *et al*. recently used classification-based ML to extract a non-intuitive structural fingerprint (named "softness"), which is strongly correlated with the probability of a particle to exhibit some rearrangement upon loading or spontaneous relaxation [31–34]. Nevertheless, due to the intrinsic complexity of the ML model, our understanding of how glass dynamics is controlled by its static

structure is still limited [25,31]. Specifically, although a few studies revealed that more "liquid-like" local neighborhoods tend to enhance atom mobility [24,31,33], it remains elusive what types of structural features is key to determine its "liquid-like" level and therefore control atom mobility in glasses[35–38]. Moreover, as the ML approach has thus far been applied to only some simple and small glass systems that may not capture the complex chemistry of more realistic ionocovalent oxide glasses[39–42], little is known about the level of correlation between glass dynamics and its static structure in more complex real-world glasses [30–32].

Here, inspired by the softness approach [31–33], we introduce a slightly revised definition for softness (relying on logistic regression and radial features, see below)—which we recently proposed to successfully predict creep dynamics of silicate gels from their static structure [43]—and apply it to investigate ion mobility in sodium silicate glasses, an archetypal glass with relevance in various fields such as household window [1,2], display screen [16,44], magmatic rock [10], and battery electrolyte [45,46]. It is worth to mention that, by conducting million-atom molecular dynamics simulations, we pioneeringly extend the softness approach to investigate a more realistic and larger glass system than ever before ($< 10^4$ atoms) [32,35,39]. Indeed, we find that the Na atom mobility is largely encoded in its initial softness, where softer Na atom exhibits higher mobility. Importantly, the use of logistic regression allows us to interpret the machine-learned softness metric. By decoding the softness, we conclude that the sodium ion mobility is highly controlled by the local density of defect oxygen neighbors that are located in between the first and second coordination shells.

10.2 Methods

To establish our conclusions, we simulate the spontaneous relaxation process of a $(Na_2O)_{30}(SiO_2)_{70}$ glass. First, we construct a large configuration that contains 1 million atoms (i.e., 205,800 Na atoms) in a cubic box with periodic boundary conditions and a side length of 241 Å, in agreement with the experimental density (2.466 g/cm^3) [47]. The interatomic potential adopted herein is the well-established Teter potential [14,47,48], which has been demonstrated to offer an accurate description of various structural, dynamical, and thermodynamical properties of silicate glasses [10,12,47,49]. The *NVT* ensemble is applied in all simulations, and the timestep is fixed as 1 fs. The system is initially melted at 4000 K for 100 picoseconds. The glass is prepared by melt quenching from 4000 K to 700 K with a cooling rate of 1 K/ps. Finally, we conduct the relaxation simulation of the glass at 700 K for 50 picoseconds and track the location of atoms (Na) over time. This temperature is large enough to activate the motion of Na atom but simultaneously low enough to ensure that Si and O network-forming atoms remain largely immobile within the time of the simulation. All simulations are performed by using the LAMMPS code [50].

10.3 Results and Discussion

Figure 10-1 shows the distribution of the Na atoms' displacement D at the end of the relaxation simulation. D is calculated as the distance between the atom's initial and final position during the relaxation. Notably, the distribution profile features two peaks associated with two atom ensembles—namely, the ensemble of immobile ($D < D_0$) and mobile ($D \geq D_0$) Na atoms, respectively, wherein D_0 is the threshold displacement that distinguishably separate the two peaks (herein, $D_0 = 2$ Å, i.e., the local minimum between the two peaks), and the two ensembles represent the populations of Na atoms that are (i) simply vibrating while being remaining trapped in their

local pocket and (ii) Na atoms that have jumped to another pocket during the time of the simulation, respectively [13]. In the following, we use this threshold D_0 to classify Na atoms as immobile (low displacement) or mobile (high displacement). Based on this threshold, about 39.7% Na atoms are classified as mobile during the relaxation. Note, however, that the following analysis is largely insensitive to small variations of the selected threshold and does not significantly depend on the arbitrary choice of this threshold displacement.

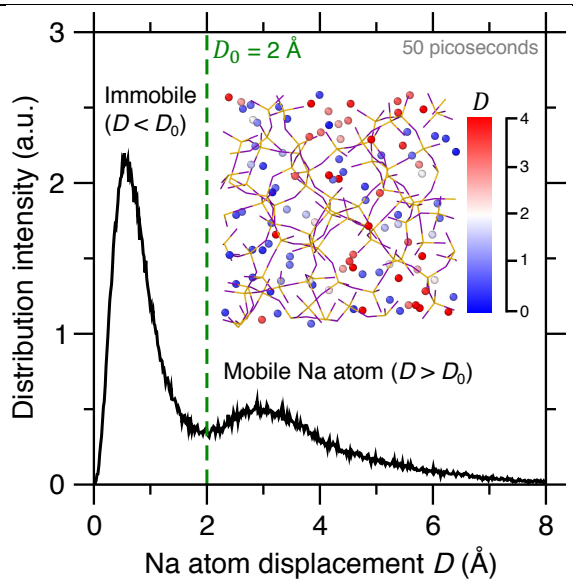


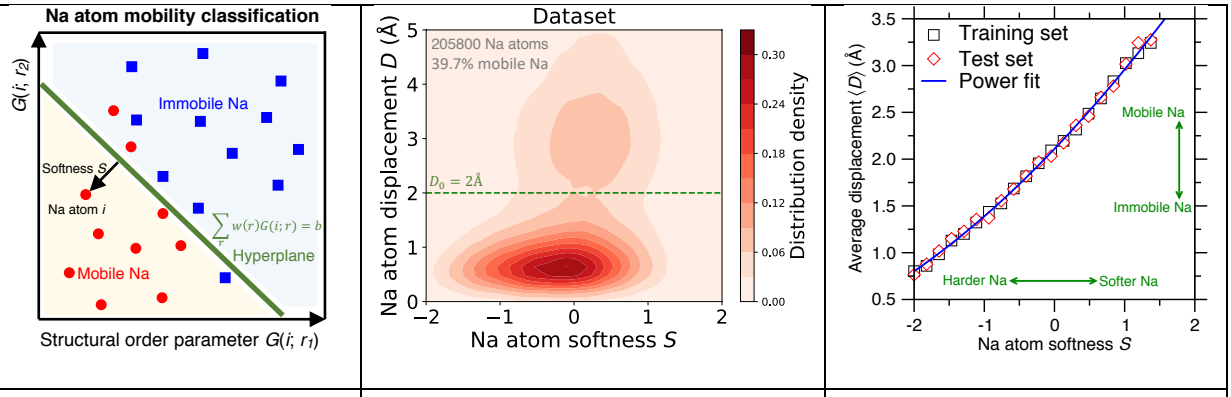
Figure 10-1: Distribution of the Na atoms' displacement D in a $(Na_2O)_{30}(SiO_2)_{70}$ glass at the end of the relaxation simulation. The system contains 205,800 Na atoms and is relaxed at a constant temperature (700 K) and volume for 50 picoseconds. The green dash refers to a selected threshold displacement D_0 = 2 Å that discriminates mobile Na atoms from immobile Na atoms. The inset is a colormap of the Na atoms' displacement in the bonded silicate network.

We now investigate whether the propensity for an Na atom to be mobile or immobile (i.e., a dynamic property) could be in some way encoded in its initial static structure. To this end,

following the example of the softness approach [31–33], we construct by machine learning a structural quantity that is correlated with the mobility of Na atoms during the relaxation process. Briefly, based on the present simulation, we first build a dataset that contains 205,800 Na atoms from the large $(Na_2O)_{30}(SiO_2)_{70}$ glass configuration simulated herein, where 70% of the Na atoms serve as training set. Note that the size of the training set has been proved to be large enough to eliminate the risk of sample deficiency for the ML model training. Then, all Na atoms are labeled as mobile ($D \geq D_0$) or immobile ($D < D_0$) by comparing their displacement D with the threshold displacement D_0 at the end of the relaxation simulation. We then train a classifier to identify an optimal classification hyperplane that separates mobile from immobile Na atoms in a standardized N_r-dimensional classification space, as illustrated in Fig. 10-2a. Here, the N_r input features (i.e., the classification space) of the classifier is constructed by computing (based on the initial static structure, before the relaxation simulation) a series of N_r radial order parameters $G(i; r)$ that describe the local oxygen density of each Na atom i at different distances r [33,43]:

$$G(i;r) = \sum_j e^{-(R_{ij}-r)^2/L^2} \qquad \textbf{Eq. (10-1)}$$

where j refers to the neighbor O atom of Na atom i within a cutoff distance R_G (here, R_G = 8 Å [33,51]), R_{ij} is the distance between the atom i and j, and L is the standard deviation of the Gaussian functions centered around r (here, L = 0.2 Å [33,51]). Overall, we calculate for each Na atom a series of $G(i; r)$ ranging from r = 1.6 Å to r = 6.4 Å with an increment of 0.3 Å [33,51], and the ensemble of these metrics offers an unbiased fingerprint of the local radial order around each Na atom.



(a) **(b)** **(c)**

Figure 10-2: (a) Schematic of the classification model used to separate mobile Na atoms (red circle) from immobile Na atoms (blue square) using a classification hyperplane (green line). The input features are constructed by a series of N_r structural order parameters $G(i; r)$ that describe the local oxygen density of each Na atom i at different distances r (see Eq. (10-1)). For illustration purpose, here, two input features associated with the distances r_1 = 2.36 Å and r_2 = 4.68 Å (i.e., the average distance of the first and second coordination shell, respectively) are selected to represent the N_r-dimensional feature space. The hyperplane is identified by logistic regression. **(b)** Distribution density of the Na atoms' final displacement D and initial softness S. The softness S is defined as the orthogonal distance between the atom and the hyperplane in classification space (see panel (a)). Mobile and immobile atoms correspond to positive and negative S, respectively. The dataset contains 205,800 Na atoms from a large $(Na_2O)_{30}(SiO_2)_{70}$ configuration with 39.7% mobile Na ($D \geq D_0$) and is randomly divided into the training (70%) and test sets (30%). **(c)** Final average Na atom displacement ‹D› of the training and test sets as a function of their initial softness S. The blue line is a power fit to guide the eye.

Unlike the original softness approach that uses as inputs both radial and angular order features [31–33], we here solely focus on radial features capturing 2-body correlations around each Na atom. This is key to ensure that the new softness metric remains highly interpretable (see below). Note that, since the Na–O interaction is nondirectional, incorporating angular 3-body order parameters does not notably increase the classification accuracy, in agreement with previous studies in Lennard-Jones systems [31,34]. In that regard, limiting the number of input features also ensures that the model does not become overfitted. Moreover, as an alternative to the support vector machine-based classifying technique adopted by the original softness approach [31–33], we use logistic regression to build the classifier [52], which offers great model simplicity, accuracy, and interpretability [43,52]. Indeed, logistic regression directly provides the probability of a given atom to be mobile or immobile. In addition, it embeds regularization to limit the risk of overfitting. Importantly, the classification hyperplane determined by logistic regression is linear, which makes it possible to easily assess the importance of each feature. We also expect that the linear nature of the hyperplane is key to enhance the extrapolability of the classification model.

We now analyze the outcome of the classification. For each Na atom, we extract a synthetic, local structural quantity "softness" from the classifier, where the softness S is defined as the orthogonal distance between the atom and the hyperplane in classification space (see Fig. 10-2a), and mobile and immobile atoms correspond to positive and negative values of S, respectively. Figure 10-2b shows the distribution density of the Na atoms' final displacement D and initial softness S. We find that the softness sign ($S > 0$ or $S < 0$) can properly separate mobile ($D \geq D_0$) and immobile ($D < D_0$) Na atoms with a decent classification accuracy of ~63% for both the training and test sets. Note that the accuracy remains limited as a large number of soft Na atoms ($S > 0$) remain trapped in the "cage effect" ($D < D_0$) under the fairly low relaxation temperature

selected herein [13]. It should be pointed out that, in contrast, at elevated temperature, the static structure tends to lose its predictivity of the long-time glass dynamics as the system quickly loses the memory of its initial structure [25]. Further, Figure 10-2c shows the final average Na atom displacement ‹*D*› of both the training and test sets as a function of their initial softness *S*. Interestingly, we find that the magnitude of Na atom displacement features a power-law dependance on softness, where softer Na atoms exhibits larger displacement during the relaxation, and vice versa. This power-law correlation is likely rooted in an intimate link between Na atom softness and the energy barrier for the atom to rearrange during relaxation (see below), which echoes recent studies that reveal a generic power-law relationship between particle displacement and the associated energy barrier to overcome in disordered materials [53,54].

For Na atoms belonging to the test set, it is notable that the degree of correlation remains high between their softness and dynamics. Figure 10-3a shows a snapshot of the predicted Na atom softness in a new 3000-atom $(Na_2O)_{30}(SiO_2)_{70}$ glass that is simulated herein as a fully independent test set. The distribution of softness (both for all Na atoms and for mobile Na atoms in the glass) is provided in Fig. 10-3b. We find that the classification accuracy is satisfactory as ~64% of the mobile Na atoms indeed exhibit a positive softness ($S > 0$). Further, we calculate the probability $P_R(S)$ of a Na atom to rearrange ($D \geq D_0$) as a function of its initial softness *S* (Fig. 10-3c). Interestingly, we find that $P_R(S)$ exhibits an exponential dependance on *S*, as following an activated process [55,56]:

$$P_R(S) = P_0 \exp(\beta S) \qquad \textbf{Eq. (10-2)}$$

where P_0 and β are some fitting parameters. In accordance with the power-law correlation between the Na atoms' final displacement *D* and initial softness *S* (see Fig. 10-2c), this exponential correlation between $P_R(S)$ and *S* suggests that the structural quantity *S* is closely related to (and

might be indicative of) the energy barrier associated with Na atom rearrangement [31,42,57], as suggested in a recent study of the creep dynamics of gels [43]. Note that softness is calculated based on the sole knowledge of the initial structure, whereas the Na atom displacement is computed at the end of the relaxation simulation. The high degree of correlation between initial softness and final displacement clearly illustrates the intimate link between glass dynamics and its initial static structure.

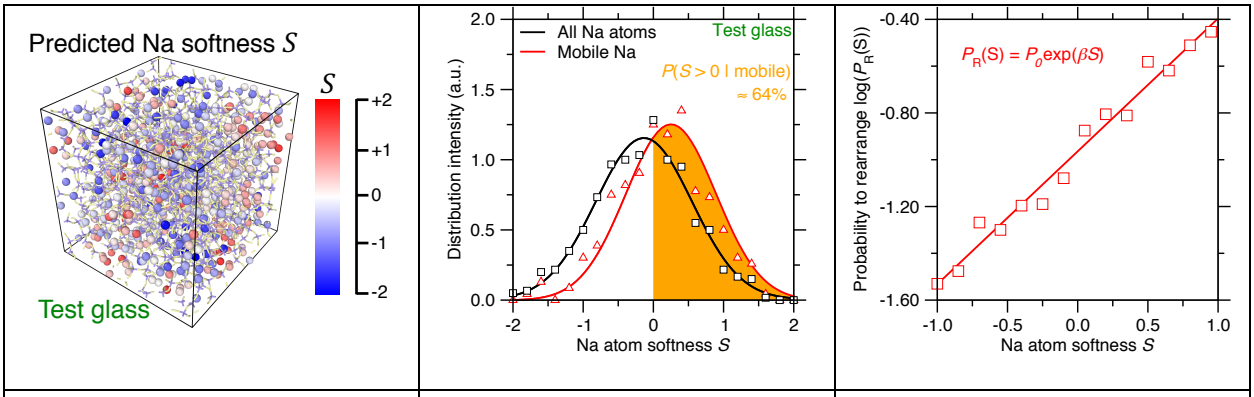


Figure 10-3: (a) Snapshot of the predicted Na atom softness S for a new, independent test $(Na_2O)_{30}(SiO_2)_{70}$ glass. The system contains 600 Na atoms as a test set. **(b)** Distribution of the softness of all Na atoms (black) and mobile Na atoms (red) in the glass. The orange area represents the properly predicted soft Na atoms ($S > 0$) within the mobile Na atoms. **(c)** Logarithm of the probability $\log(P_R(S))$ of a Na atom to rearrange ($D \geq D_0$) as a function of its initial softness S. The red line is an exponential fit following Eq. (10-2).

Finally, we discuss the structural interpretation of the machine-learned softness metric. Indeed, the hyperplane created by logistic regression can be expressed as a linear equation of each of the features as (see Fig. 10-2a):

$$\Sigma_r w(r) G(i; r) = b \qquad \textbf{Eq. (10-3)}$$

wherein $w(r)$ and b are the coefficients and the bias of the logistic regression model, respectively. Note that, all input features $G(i; r)$ have been standardized (before training) so that the coefficients are directly indicative of the relative importance of each feature in the classification. Namely, a large absolute value for $w(r)$ refers to a fairly orthogonal hyperplane to the axis associated with the corresponding feature $G(i; r)$. In addition, the positive and negative sign of the coefficients $w(r)$ is informative as it indicates that increasing values of the feature $G(i; r)$ tend to result in increased and decreased softness values, respectively.

Figure 10-4a shows the coefficients $w(r)$ of the logistic regression classifier as a function of the distance r, wherein the absolute value of $w(r)$ denotes how influential the feature $G(i; r)$ is on determining the atom softness. We find that the most influential feature is associated with the distance r_m that corresponds to the region that is located between the peak positions r_1 and r_2 of, respectively, the 1st and 2nd coordination shells of the Na–O partial pair distribution function $g_{\text{Na–O}}(r)$ (see the upper panel of Fig. 10-4). Note that $w(r)$ is negative at the distance r_m as well as at all other distances. Although the absolute values of $w(r)$ at other distances are smaller and approaches zero when r is larger than r_2, we notice that the features $G(i; r)$ that are associated with distances r that are close to r_m (in between r_1 and r_2) contribute significantly more than other features to determine softness (see the grey window in Fig. 10-4). Namely, a "defect" oxygen neighbor located in this extent of distances (i.e., in between the first and second coordination shells) tends to greatly reduce the Na atom's mobility (see blue particle in Fig. 10-4a), although all oxygen neighbors within the first two coordination shells synergically reduce the mobility of the central Na atom. The key role played by the "defect" oxygen neighbors at the inter-shell region is to occupy those potential empty jumping sites around the central Na atom so as to block the Na atom's motion within the displacement threshold ($D < D_0$).

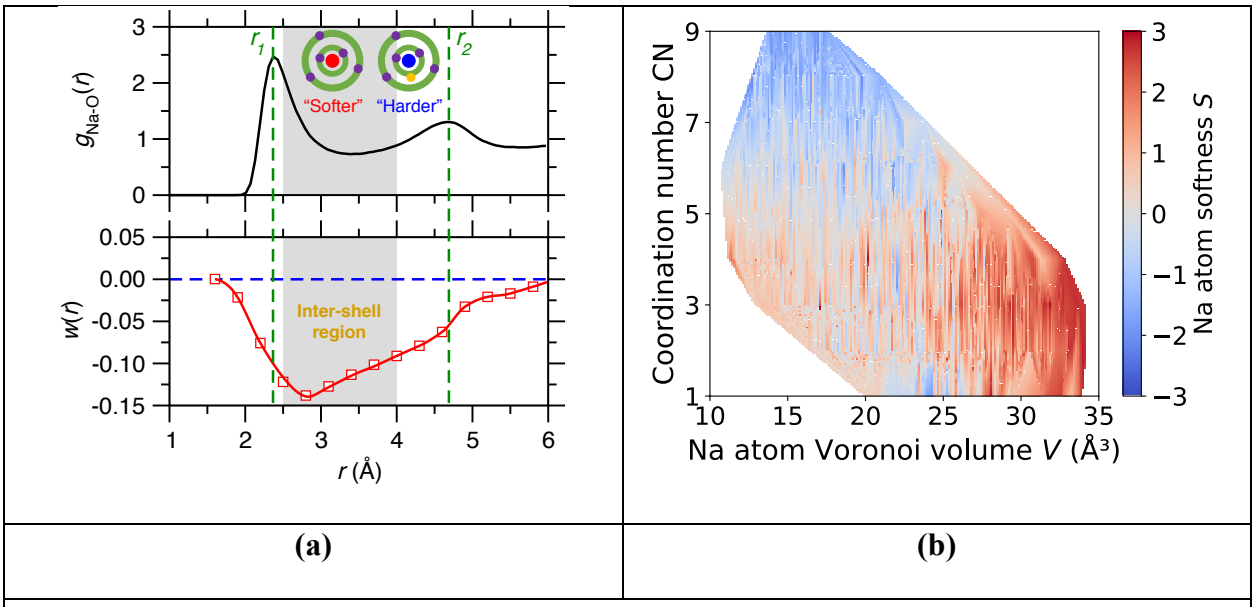


Figure 10-4: (a) Weight coefficient $w(r)$ of the classification hyperplane (see Fig. 10-2a) at different distances r. The red line is to guide the eyes. The partial Na–O pair distribution function $g_{Na–O}(r)$ of the glass is added in the top panel as reference. The distance r_1 and r_2 are associated with the peak position of the 1st and 2nd coordination shell of $g_{Na–O}(r)$, respectively. The grey window indicates the range of large weights. The inset illustrates the local oxygen (purple sphere) environments around (i) a soft Na atom (red sphere) and (ii) a "harder" Na atom (blue sphere) with an extra O atom (gold sphere) in between the 1st and 2nd coordination shells (green halo). **(b)** Na atom softness S as a function of their Voronoi volume V and coordination number CN. The color coding is based on a linear interpolation between the datapoints in the Na atom dataset.

These results are consistent with free volume theory [13,14,48]. Indeed, closed-packed structures with a large number of oxygen neighbors are associated with low local free volume, wherein the central Na atom exhibits very limited mobility. In contrast, more loosely-packed structures exhibiting less oxygen neighbors tend to show more potential empty jumping sites

around them, which promotes Na atom mobility (see red particle in Fig. 10-4a). Further, Figure 10-4b illustrates the dependance of Na atoms' softness S on their Voronoi volume V and coordination number CN. Overall, larger CN and smaller V values tend to favor smaller softness. However, we nevertheless observe that softness is a complex, nonmonotonic function of CN and V. Indeed, we find that the classifier trained by the sole knowledge of the Na atom Voronoi volume offers a very limited accuracy of ~50% as compared to that offered by the softness metric (~63% accuracy). Similarly, training a classifier based on the sole knowledge of the Na atom coordination number yields an accuracy of ~52%, wherein both low and high-coordination atoms are very likely to be classified as soft. This indicates that, although they offer an intuitive interpretation of the origin of Na mobility, the coordination number and Voronoi volume metrics do not fully capture the propensity of Na to reorganize. This exemplifies the benefit of using an unbiased machine learning approach to build the set of input features, since intuitive structural features show only limited correlation with dynamical properties.

10.4 Conclusions

Overall, these results highlight the ability of machine learning to analyze large amounts of complex data and decode previously hidden correlations—here, between the dynamics of a glass and its initial static structure. It is notable that our approach allows us to predict the dynamics of a realistic, complex oxide glass based on the sole knowledge of the machine-learned softness metric. The interpretation of the softness metric defined herein (see Fig. 10-4a) suggests that the mobility of sodium atoms in silicate glasses is strongly anticorrelated with the local density of defect oxygen neighbors that are located in between the 1st and 2nd coordination shells. Machine learning

therefore offers a promising route to decode the complex relationship between structure and properties in disordered, out-of-equilibrium phases.

10.5 References

[1] J.D. Musgraves, J. Hu, L. Calvez, eds., Springer Handbook of Glass, Springer International Publishing, Cham, 2019. https://doi.org/10.1007/978-3-319-93728-1.

[2] J.C. Mauro, E.D. Zanotto, Two Centuries of Glass Research: Historical Trends, Current Status, and Grand Challenges for the Future, International Journal of Applied Glass Science. 5 (2014) 313–327. https://doi.org/10.1111/ijag.12087.

[3] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[4] R.C. Welch, J.R. Smith, M. Potuzak, X. Guo, B.F. Bowden, T.J. Kiczenski, D.C. Allan, E.A. King, A.J. Ellison, J.C. Mauro, Dynamics of Glass Relaxation at Room Temperature, Phys. Rev. Lett. 110 (2013) 265901. https://doi.org/10.1103/PhysRevLett.110.265901.

[5] Y. Yu, M. Wang, D. Zhang, B. Wang, G. Sant, M. Bauchy, Stretched Exponential Relaxation of Glasses at Low Temperature, Physical Review Letters. 115 (2015). https://doi.org/10.1103/PhysRevLett.115.165901.

[6] Y. Yu, M. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Bauchy, Thermometer Effect: Origin of the Mixed Alkali Effect in Glass Relaxation, Phys. Rev. Lett. 119 (2017) 095501. https://doi.org/10.1103/PhysRevLett.119.095501.

[7] C. Massobrio, ed., Molecular dynamics simulations of disordered materials: from network glasses to phase-change memory allyos, Springer, Cham Heidelberg, 2015.

[8] M. Bauchy, Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: A review, Computational Materials Science. 159 (2019) 95–102. https://doi.org/10.1016/j.commatsci.2018.12.004.

[9] J.C. Mauro, Y. Yue, A.J. Ellison, P.K. Gupta, D.C. Allan, Viscosity of glass-forming liquids, Proceedings of the National Academy of Sciences. 106 (2009) 19780–19784. https://doi.org/10.1073/pnas.0911705106.

[10] M. Bauchy, B. Guillot, M. Micoulaut, N. Sator, Viscosity and viscosity anomalies of model silicates and magmas: A numerical investigation, Chemical Geology. 346 (2013) 47–56. https://doi.org/10.1016/j.chemgeo.2012.08.035.

[11] S.S. Sørensen, M.B. Østergaard, M. Stepniewska, H. Johra, Y. Yue, M.M. Smedskjaer, Metal–Organic Framework Glasses Possess Higher Thermal Conductivity than Their Crystalline Counterparts, ACS Appl. Mater. Interfaces. 12 (2020) 18893–18903. https://doi.org/10.1021/acsami.0c02310.

[12] M. Bauchy, Structural, vibrational, and thermal properties of densified silicates: Insights from molecular dynamics, The Journal of Chemical Physics. 137 (2012) 044510. https://doi.org/10.1063/1.4738501.

[13] M. Bauchy, M. Micoulaut, From pockets to channels: Density-controlled diffusion in sodium silicates, Phys. Rev. B. 83 (2011) 184118. https://doi.org/10.1103/PhysRevB.83.184118.

[14] A.N. Cormack, J. Du, T.R. Zeitler, Sodium ion migration mechanisms in silicate glasses probed by molecular dynamics simulations, Journal of Non-Crystalline Solids. 323 (2003) 147–154. https://doi.org/10.1016/S0022-3093(03)00280-1.

[15] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Accelerating the Design of Functional Glasses through Modeling, Chem. Mater. 28 (2016) 4267–4277. https://doi.org/10.1021/acs.chemmater.6b01054.

[16] M.C. Onbaşlı, A. Tandia, J.C. Mauro, Mechanical and Compositional Design of High-Strength Corning Gorilla® Glass, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling: Applications: Current and Emerging Materials, Springer International Publishing, Cham, 2018: pp. 1–23. https://doi.org/10.1007/978-3-319-50257-1_100-1.

[17] E.D. Zanotto, F.A.B. Coutinho, How many non-crystalline solids can be made from all the elements of the periodic table?, Journal of Non-Crystalline Solids. 347 (2004) 285–288. https://doi.org/10.1016/j.jnoncrysol.2004.07.081.

[18] K. Binder, W. Kob, Glassy Materials and Disordered Solids: An Introduction to Their Statistical Mechanics, World Scientific, 2011.

[19] S.S. Sørensen, C.A.N. Biscio, M. Bauchy, L. Fajstrup, M.M. Smedskjaer, Revealing hidden medium-range order in amorphous materials using topological data analysis, Sci. Adv. 6 (2020) eabc2320. https://doi.org/10.1126/sciadv.abc2320.

[20] M.A. Klatt, J. Lovrić, D. Chen, S.C. Kapfer, F.M. Schaller, P.W.A. Schönhöfer, B.S. Gardiner, A.-S. Smith, G.E. Schröder-Turk, S. Torquato, Universal hidden order in amorphous cellular geometries, Nature Communications. 10 (2019) 811. https://doi.org/10.1038/s41467-019-08360-5.

[21] M. Mungan, S. Sastry, K. Dahmen, I. Regev, Networks and Hierarchies: How Amorphous Materials Learn to Remember, Phys. Rev. Lett. 123 (2019) 178002. https://doi.org/10.1103/PhysRevLett.123.178002.

[22] R.L. Jack, A.J. Dunleavy, C.P. Royall, Information-Theoretic Measurements of Coupling between Structure and Dynamics in Glass Formers, Phys. Rev. Lett. 113 (2014) 095703. https://doi.org/10.1103/PhysRevLett.113.095703.

[23] A. Widmer-Cooper, P. Harrowell, H. Fynewever, How Reproducible Are Dynamic Heterogeneities in a Supercooled Liquid?, Phys. Rev. Lett. 93 (2004) 135701. https://doi.org/10.1103/PhysRevLett.93.135701.

[24] E.D. Cubuk, S.S. Schoenholz, E. Kaxiras, A.J. Liu, Structural Properties of Defects in Glassy Liquids, J. Phys. Chem. B. 120 (2016) 6139–6146. https://doi.org/10.1021/acs.jpcb.6b02144.

[25] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E.D. Cubuk, S.S. Schoenholz, A. Obika, A.W.R. Nelson, T. Back, D. Hassabis, P. Kohli, Unveiling the predictive power of static structure in glassy systems, Nat. Phys. 16 (2020) 448–454. https://doi.org/10.1038/s41567-020-0842-8.

[26] S.J. Russell, P. Norvig, Artificial Intelligence : A Modern Approach, Malaysia; Pearson Education Limited, 2016. http://thuvienso.thanglong.edu.vn/handle/DHTL_123456789/4010.

[27] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: A review, Journal of Non-Crystalline Solids: X. 4 (2019) 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[28] R. Iten, T. Metger, H. Wilming, L. del Rio, R. Renner, Discovering Physical Concepts with Neural Networks, Phys. Rev. Lett. 124 (2020) 010508. https://doi.org/10.1103/PhysRevLett.124.010508.

[29] G. Biroli, Machine learning glasses, Nat. Phys. 16 (2020) 373–374. https://doi.org/10.1038/s41567-020-0873-1.

[30] Z. Fan, J. Ding, E. Ma, Machine learning bridges local static structure with multiple properties in metallic glasses, Materials Today. 40 (2020) 48–62. https://doi.org/10.1016/j.mattod.2020.05.021.

[31] S.S. Schoenholz, E.D. Cubuk, D.M. Sussman, E. Kaxiras, A.J. Liu, A structural approach to relaxation in glassy liquids, Nature Physics. 12 (2016) 469–471. https://doi.org/10.1038/nphys3644.

[32] E.D. Cubuk, R.J.S. Ivancic, S.S. Schoenholz, D.J. Strickland, A. Basu, Z.S. Davidson, J. Fontaine, J.L. Hor, Y.-R. Huang, Y. Jiang, N.C. Keim, K.D. Koshigan, J.A. Lefever, T. Liu, X.-G. Ma, D.J. Magagnosc, E. Morrow, C.P. Ortiz, J.M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K.N. Nordstrom, P.E. Arratia, R.W. Carpick, D.J. Durian, Z. Fakhraai, D.J. Jerolmack, D. Lee, J. Li, R. Riggleman, K.T. Turner, A.G. Yodh, D.S. Gianola, A.J. Liu, Structure-property relationships from universal signatures of plasticity in disordered solids, Science. 358 (2017) 1033–1037. https://doi.org/10.1126/science.aai8830.

[33] E.D. Cubuk, S.S. Schoenholz, J.M. Rieser, B.D. Malone, J. Rottler, D.J. Durian, E. Kaxiras, A.J. Liu, Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods, Physical Review Letters. 114 (2015). https://doi.org/10.1103/PhysRevLett.114.108001.

[34] D.M. Sussman, S.S. Schoenholz, E.D. Cubuk, A.J. Liu, Disconnecting structure and dynamics in glassy thin films, PNAS. 114 (2017) 10601–10605. https://doi.org/10.1073/pnas.1703927114.

[35] Q. Wang, A. Jain, A transferable machine-learning framework linking interstice distribution and plastic heterogeneity in metallic glasses, Nat Commun. 10 (2019) 1–11. https://doi.org/10.1038/s41467-019-13511-9.

[36] H. Tanaka, H. Tong, R. Shi, J. Russo, Revealing key structural features hidden in liquids and glasses, Nature Reviews Physics. 1 (2019) 333–348. https://doi.org/10.1038/s42254-019-0053-3.

[37] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, L. Filion, Autonomously revealing hidden local structures in supercooled liquids, Nature Communications. 11 (2020) 5479. https://doi.org/10.1038/s41467-020-19286-8.

[38] Q. Wang, J. Ding, L. Zhang, E. Podryabinkin, A. Shapeev, E. Ma, Predicting the propensity for thermally activated β events in metallic glasses via interpretable machine learning, Npj Comput Mater. 6 (2020) 194. https://doi.org/10.1038/s41524-020-00467-4.

[39] E.D. Cubuk, A.J. Liu, E. Kaxiras, S.S. Schoenholz, Unifying framework for strong and fragile liquids via machine learning: a study of liquid silica, ArXiv:2008.09681 [Cond-Mat]. (2020). http://arxiv.org/abs/2008.09681 (accessed August 29, 2020).

[40] J. Du, Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 157–180. https://doi.org/10.1007/978-3-319-15675-0_7.

[41] L. Huang, J. Kieffer, Challenges in Modeling Mixed Ionic-Covalent Glass Formers, in: C. Massobrio, J. Du, M. Bernasconi, P.S. Salmon (Eds.), Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys, Springer International Publishing, Cham, 2015: pp. 87–112. https://doi.org/10.1007/978-3-319-15675-0_4.

[42] X. Ma, Z.S. Davidson, T. Still, R.J.S. Ivancic, S.S. Schoenholz, A.J. Liu, A.G. Yodh, Heterogeneous Activation, Local Structure, and Softness in Supercooled Colloidal Liquids, Physical Review Letters. 122 (2019). https://doi.org/10.1103/PhysRevLett.122.028001.

[43] H. Liu, S. Xiao, L. Tang, E. Bao, E. Li, C. Yang, Z. Zhao, G. Sant, M.M. Smedskjaer, L. Guo, M. Bauchy, Predicting the early-stage creep dynamics of gels from their static structure by machine learning, Acta Materialia. 210 (2021) 116817. https://doi.org/10.1016/j.actamat.2021.116817.

[44] M.C. Onbaşlı, J.C. Mauro, Modeling the Relaxation Behavior of Glasses for Display Applications, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling, Springer International Publishing, Cham, 2019: pp. 1–19. https://doi.org/10.1007/978-3-319-50257-1_99-1.

[45] M.H. Braga, A.J. Murchison, J.A. Ferreira, P. Singh, J.B. Goodenough, Glass-amorphous alkali-ion solid electrolytes and their performance in symmetrical cells, Energy Environ. Sci. 9 (2016) 948–954. https://doi.org/10.1039/C5EE02924D.

[46] Z.A. Grady, C.J. Wilkinson, C.A. Randall, J.C. Mauro, Emerging Role of Non-crystalline Electrolytes in Solid-State Battery Research, Front. Energy Res. 8 (2020). https://doi.org/10.3389/fenrg.2020.00218.

[47] J. Du, A.N. Cormack, The medium range structure of sodium silicate glasses: a molecular dynamics simulation, Journal of Non-Crystalline Solids. 349 (2004) 66–79. https://doi.org/10.1016/j.jnoncrysol.2004.08.264.

[48] A.N. Cormack, J. Du, T.R. Zeitler, Alkali ion migration mechanisms in silicate glasses probed by molecular dynamics simulations, Phys. Chem. Chem. Phys. 4 (2002) 3193–3197. https://doi.org/10.1039/B201721K.

[49] J. Du, L.R. Corrales, Compositional dependence of the first sharp diffraction peaks in alkali silicate glasses: A molecular dynamics study, Journal of Non-Crystalline Solids. 352 (2006) 3255–3269. https://doi.org/10.1016/j.jnoncrysol.2006.05.025.

[50] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[51] S. Ispas, M. Benoit, P. Jund, R. Jullien, Structural properties of glassy and liquid sodium tetrasilicate: comparison between ab initio and classical molecular dynamics simulations, Journal of Non-Crystalline Solids. 307–310 (2002) 946–955. https://doi.org/10.1016/S0022-3093(02)01549-1.

[52] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[53] L. Tang, G. Ma, H. Liu, W. Zhou, M. Bauchy, Bulk Metallic Glasses' Response to Oscillatory Stress Is Governed by the Topography of the Energy Landscape, J. Phys. Chem. B. (2020) acs.jpcb.0c08794. https://doi.org/10.1021/acs.jpcb.0c08794.

[54] L. Tang, H. Liu, G. Ma, T. Du, N. Mousseau, W. Zhou, M. Bauchy, The Energy Landscape Governs Ductility in Disordered Materials, Mater. Horiz. (2021). https://doi.org/10.1039/D0MH00980F.

[55] A. Nicolas, E.E. Ferrero, K. Martens, J.-L. Barrat, Deformation and flow of amorphous solids: Insights from elastoplastic models, Rev. Mod. Phys. 90 (2018) 045006. https://doi.org/10.1103/RevModPhys.90.045006.

[56] S.R. Elliott, F.E.G. Henn, Application of the Anderson-Stuart model to the AC conduction of ionically conducting materials, Journal of Non-Crystalline Solids. 116 (1990) 179–190. https://doi.org/10.1016/0022-3093(90)90691-E.

[57] R. Freitas, E.J. Reed, Uncovering the effects of interface-induced ordering of liquid on crystal growth using machine learning, Nat Commun. 11 (2020) 1–10. https://doi.org/10.1038/s41467-020-16892-4.

Chapter 11. Predicting the Early-Stage Creep Dynamics of Gels from Their Static Structure by Machine Learning

11.1 Introduction

When subjected to a sustained load, disordered solids (e.g., glasses, granular materials, or gels) tend to exhibit delayed, time-dependent creep deformations [1–3]. Although creep can occur in many types of materials (e.g., metals or ceramics, primarily at high temperature [4,5]), it is especially pronounced in soft matter, e.g., colloidal gels [3,6,7]. In that regard, the viscoplastic deformation of calcium–silicate–hydrate gels (C–S–H, the binding phase of concrete) under constant load plays a key role in the built environment since it is responsible for concrete's creep [7–10].

Despite the important, often detrimental role of creep in colloidal gels, its nanoscale origin, driving force, and mechanism remain debated [3,6,8]. In particular, it remains unclear whether the propensity of a disordered solid to creep could in some way be encoded in its static, unloaded structure [11]. This question is a manifestation of a more general gap in our understanding of how structure controls dynamics in disordered phases [6,11,12]. Indeed, due to the complex, disordered structure of glasses or gels [13,14], pinpointing which structural features govern dynamics is essentially a "needle-in-a-haystack" problem [15–17], since intuitive structural metrics (e.g., local packing or coordination number) are often only weakly correlated with dynamics [18,19].

Owing to its ability to discover relevant patterns in complex, multidimensional data, machine learning (ML) offers a new opportunity to revisit the nature of the linkages between structure and dynamics in disordered phases—without the need for any prerequisite intuition regarding which structural feature(s) could be influential [20,21]. In particular, Cubuk *et al.* recently extracted by ML a non-intuitive structural fingerprint (named "softness"), which is

strongly correlated with the probability of a particle to exhibit some rearrangement upon loading or spontaneous relaxation [12,15,22–26]. However, although softness has been shown to be correlated with near-future particle rearrangements, it has thus far been unable to offer insights into the long-time dynamics of disordered phases [11]. This has prevented the use of this approach to study creep, which can extend over several years [9].

Here, inspired by this softness approach, we introduce a slightly revised definition for softness (relying on a linear logistic regression model and radial features) and use this machine-learned structural fingerprint to interrogate the existence of a causal link between structure and long-time creep. This approach reveals that the propensity of a colloidal gel to creep is encoded in its instantaneous, static structure. Importantly, we find that the softness metric captures the effective average energy barrier that the particles need to overcome to rearrange during creep—which suggests that the softness metric offers a structural fingerprint of the topography of the energy landscape. Finally, the use of linear logistic regression allows us to offer a structural interpretation of this machine-learned predictor.

11.2 Methods

11.2.1 Archetypical gel model

To establish our conclusions, we simulate an archetypical mesoscale model of a colloidal C–S–H gel [10,27,28]. The model gel is comprised of an ensemble of monodisperse spherical particles of 5 nm. The interaction between particles is described by a generalized Lennard-Jones potential with a minimum at distance $\sigma = \sqrt[14]{2} \times 5$ nm, which corresponds to the effective particle diameter. In detail, the potential is descried as [27]:

$$U_{ij}(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma_0}{r_{ij}}\right)^{2\alpha} - \left(\frac{\sigma_0}{r_{ij}}\right)^{\alpha}\right] \qquad \textbf{Eq. (11-1)}$$

where σ_0 is the particle diameter (5 nm here), α a parameter that controls the narrowness of the potential well (here, α = 14), r_{ij} the distance between the centers of a pair of particles i and j, and ε the depth of the potential energy well and $\varepsilon = A_0{\sigma_0}^3$, where $A_0 = kE$ is a prefactor that is proportional to the Young's modulus E of a bulk C–S–H grain (here, $E = 63.6$ GPa and $k = 0.002324$). The potential defined in Eq. (11-1) exhibits a minimum at $r_m = \sqrt[\alpha]{2}\sigma_0$ so that the effective diameter of a particle i is here defined as $\sigma = \sqrt[\alpha]{2}\sigma_0$. This model has been extensively studied and has been shown to offer a realistic description of the structure and mechanical properties of C–S–H gels [8,10,27–31].

11.2.2 Preparation of the gel configurations

The gel configurations are generated by grand canonical Monte Carlo (GCMC) simulations, wherein particles are iteratively inserted until saturation into an initially empty cubic box of 600 Å length with periodic boundary. Each GCMC step comprises of 5 attempts of particle insertions or deletions, followed by 500 attempts of random displacement of an existing particle. The temperature is fixed at T = 300 K and the excess chemical potential, controlling the probability of acceptance of the insertion attempts, is set to $-2k_BT$. In detail, the probability of acceptance of the attempt is given by $\min\{1,\ \exp[-(\Delta U - \mu\delta)/k_{\mathrm{B}}T]\}$, where k_{B} is the Boltzmann constant, T the temperature, ΔU the variation in potential energy caused by the trial move, μ the chemical potential (fixed at $-2k_BT$ based on Refs. [8,10,27–31], which ensures the formation of a realistic packed final structure within a reasonable simulation time), and δ the variation in the number of C–S–H grains [8,28,29]. The saturated configurations are then relaxed by molecular dynamics (MD) simulations with a timestep of 50 fs in the isothermal-isobaric (*NPT*) ensemble at 300 K and zero stress for 50 ns to release the macroscopic tensile pressure formed during the GCMC simulation.

Finally, the configurations are subjected to an energy minimization to reach their inherent structure. Note that the GCMC ensemble adopted herein aims to mimic the precipitation process of colloidal C–S–H gels and has been shown to offer a realistic description of the structure and packing density of disordered C–S–H gels [8,10,27–31]. Based on the system size considered herein, we typically get a number of particles $n_p \approx 1700$ at saturation, which corresponds to a packing density $\varphi \approx 0.63$. Using this methodology, we simulate 10 independent configurations for statistical averaging. All simulations are performed by using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) code [32].

11.2.3 Accelerated creep simulations

Since a direct MD simulation of the creep of colloidal gels is out of reach considering its extended timescale, we adopt herein a recently introduced accelerated simulation technique [33,34], which has been shown to successfully model the creep of disordered phases [2,8]. This accelerated method relies on the application of small stress perturbations to accelerate relaxation [2,33,34]. In detail, we first apply a constant, sustained shear stress τ_0 (here τ_0 = 100 MPa) to induce a creep response within the gel. The creep response of the gel is simply a linear function of τ_0, as long as τ_0 remains lower than the yield stress of the gel [8]. Small, cyclic shear stress perturbations $\pm\Delta\tau$ (here $\Delta\tau$ = 30 MPa) are then applied to accelerate the creep dynamics [8]. At each stress cycle, a constant-stress minimization of the energy is performed, wherein the system can adjust its shape and volume in order to reach the target stress. Effectively, this method mimics the accelerated relaxation exhibited by granular materials when subjected to vibrations [35]. The resulting acceleration of the dynamics arises from the fact that each stress perturbation slightly deforms the local energy landscape, which, in turn, can reduce the height of some energy barriers

that are locally accessible to the particles. This allows the system to jump over these barriers, thereby reaching a new energy basin within the landscape in an accelerated fashion [2,33,36]. Such particle reorganizations make it possible for the gel to exhibit some macroscopic viscoplastic deformation (i.e., creep) in order to accommodate the external sustained shear stress. Note that the average stress τ_0 of 100 MPa used herein is notably lower than the yield stress of the system (~600 MPa), so that no macroscopic flow of particles is observed. In addition, the stress perturbation amplitude $\Delta\tau$ (± 30 MPa) is chosen to be large enough to accelerate the creep simulation, but low enough to avoid any rejuvenation [8]. The resulting creep modulus was shown to be independent of the specific value of this stress perturbation amplitude $\Delta\tau$ [8]. Since the particle rearrangements induced by the stress perturbations are limited, the modeled gel remains in the primary creep stage (i.e., wherein the creep rate decays over time) without entering into the secondary steady-state stage (i.e., constant creep rate) or the final avalanche stage (i.e., acceleration of creep rate) [37,38]. It should also be noted that the monodisperse colloidal gel considered herein is out-of-equilibrium and tends to easily crystalize at finite temperature. Nevertheless, no crystallization is observed during the creep simulations. All simulations are performed by using the LAMMPS code [32].

11.2.4 Non-affine squared displacement of the particles

We calculate, for each particle *i*, the normalized non-affine squared displacement $D^2_{\text{min}}/\sigma^2$ at the *N*th stress perturbation cycle (here, $N = 10^6$) with respect to the initial reference configuration (here, $N_{\text{ref}} = 1$) using Eq. (11-2) [22,39]:

$$D^2_{\text{min}}/\sigma^2\,(i, N, N_{\text{ref}}) = \frac{\min_{\Lambda_i}\left\{\frac{1}{n_i}\sum_j\left[R_{ij}(N) - \Lambda_i R_{ij}(N_{\text{ref}})\right]^2\right\}}{\sigma^2} \qquad \textbf{Eq. (11-2)}$$

where R_{ij} is the distance between particle *i* and *j*, particle *j* represents the neighbor of particle *i* within a cutoff distance R_c (here, $R_c = 2\sigma$ [39]), and n_i is the total number of neighbor particles

within the range of R_c for each particle i. The quantity is minimized over choices of the local strain tensor Λ_i of particle i. Note that the quantity $\Lambda_i R_{ij}(N_{\mathrm{ref}})$ represents the distance between particles i and j after an affine deformation resulting from the application of a local strain Λ_i to the initial interparticle distance $R_{ij}(N_{\mathrm{ref}})$. This consists in computing the L^2-norm of the matrix multiplication between the local strain tensor Λ_i and the distance vector $R_{ij}(N_{\mathrm{ref}})$ between particles i and j at the N_{ref}th cycle:

$$\Lambda_i R_{ij}(N_{\mathrm{ref}}) = \left\| \begin{bmatrix} \lambda_{xx} & \lambda_{xy} & \lambda_{xz} \\ \lambda_{yx} & \lambda_{yy} & \lambda_{yz} \\ \lambda_{zx} & \lambda_{zy} & \lambda_{zz} \end{bmatrix} [r_x \quad r_y \quad r_z]^{\mathrm{T}} \right\|_2 \qquad \textbf{Eq. (11-3)}$$

where λ_{xy} is the strain component of Λ_i in the xy-axis plane and r_x is the projection of the distance vector $R_{ij}(N_{\mathrm{ref}})$ along the x-axis. The calculation of non-affine squared displacement D^2_{min} is implemented by using the OVITO software [40].

11.2.5 Average energy barrier of the particles

To explore the topography of the potential energy landscape (PEL) of the initial static gel configuration (before any stress is applied), we adopt the activation-relaxation technique nouveau (ARTn) method [41]. Starting from a local minimum of PEL, the ARTn algorithm systematically searches for the saddle points and transition pathways that are accessible from this minimum. This allows us to compute the distribution of the energy barriers (i.e., difference of energy between the saddle point and the original local minimum) that are locally accessible to each particle. In detail, starting from the initial gel configuration (located in a local minimum of the PEL), the target particle and the first-coordination neighbors thereof are first activated with a random displacement so as to identify a direction of negative curvature that denotes the presence of a nearby saddle point within the energy landscape. The activated system is then relaxed toward the saddle point by

following the direction of the negative energy curvature until the curvature is smaller than a given threshold ($0.1\,\varepsilon/\sigma^2$ herein). Finally, we compute the energy barrier associated with the target particle's rearrangement by subtracting the energy of the saddle point with the initial minimum energy. As such, the ARTn method restricts its search of particle rearrangements to those going through well-defined saddle points and, hence, focuses on the tiny fraction of the configurational space that is physically accessible to the system. To estimate the local distribution of energy barriers that are accessible to each particle, we conduct 20 independent saddle point searches for each particle— which is here found to be large enough to ensure the convergence of the energy barrier distribution [42]. Based on this analysis, we then compute the average value E_{ave} of the energy barriers that are accessible to each particle [43].

11.3 Results

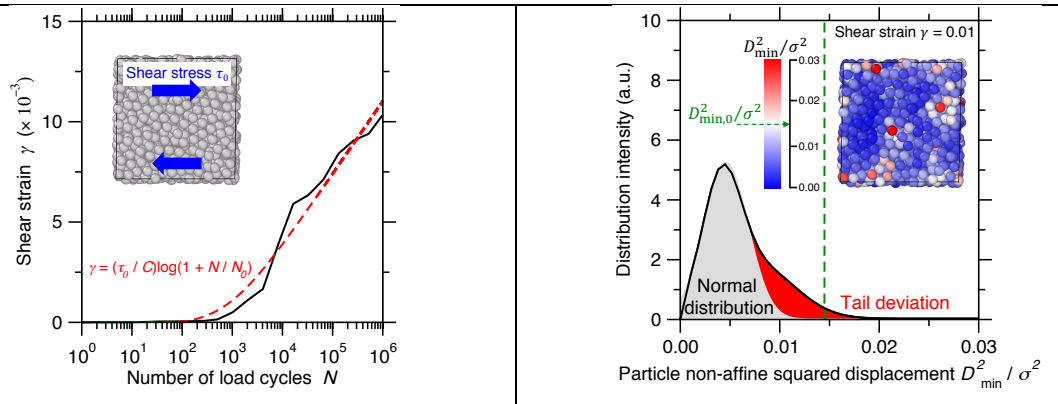
11.3.1 Long-time creep dynamics

We adopt an accelerated simulation technique based on stress perturbations to simulate the long-time creep behavior of colloidal gels subjected to a sustained shear stress τ_0 [33,34]. Details regarding the creep simulation can be found in the Methods section and in Ref. [8]. Figure 11-1a shows an example of shear strain γ evolution as a function of the number of stress perturbation cycles *N*. In agreement with previous works [8,9], our simulation predicts a logarithmic creep, which follows:

$$\gamma(N) = (\tau_0/C)\log(1 + N/N_0) \qquad \textbf{Eq. (11-4)}$$

where *C*, the creep modulus, is a material constant [8] and N_0 is a fitting parameter that is analog to a typical relaxation time [2,8]. Importantly, the creep modulus obtained from this method was shown to match experimental data obtained on C–S–H gels—which confirms that the creep

deformation induced by our accelerated method is similar to the one that would spontaneously occur over time [8].



(a) **(b)**

Figure 11-1: (a) Shear strain γ as a function of the number of stress perturbation cycles of a colloidal C–S–H gel subjected to a constant shear stress τ_0. The dashed line is a logarithmic fit following Eq. (11-4). **(b)** Distribution of the normalized non-affine squared displacement D^2_{min}/σ^2 for a shear strain γ = 1%. The red area highlights the tail of the distribution, i.e., its deviation from a Gaussian distribution (in grey). The inset shows the corresponding gel configuration, wherein the color of the particles denotes their D^2_{min}/σ^2 value. The green dash line indicates the threshold ($D^2_{min,0}/\sigma^2$) that is used herein to discriminate immobile (low displacement) from mobile (high displacement) particles.

Since the present simulation successfully reproduces the macroscopic creep of C–S–H gels, we now further analyze the simulated trajectories to explore the particle-scale mechanism of creep—which is typically hidden from conventional experiments [8,9]. To this end, we compute the normalized non-affine squared displacement D^2_{min}/σ^2 of each particle [22,39]—a metric that has been widely used to identify particle reorganizations under stress [12,44,45]. Details regarding the calculation of D^2_{min}/σ^2 can be found in the Methods section. Figure 11-1b shows the

distribution of the D^2_{min}/σ^2 values for a shear strain γ = 1%. We observe that the distribution is centered around a low displacement value (i.e., D^2_{min}/σ^2 = 0.005), which corresponds to the population of particles that do not exhibit significant reorganizations upon creep. However, we note that the distribution also exhibits a long tail toward large displacement values, suggesting that a few select particles feature some significant reorganizations. In the following, we use a threshold normalized non-affine squared displacement of $D^2_{min,0}/\sigma^2$ = 0.014 (corresponding to a displacement of about 12% of σ) to classify particles as immobile (low displacement) or mobile (high displacement). Based on this threshold, about 7% of the particles are classified as mobile during the creep process. Note, however, that the following analysis does not significantly depend on the arbitrary choice of this threshold displacement.

11.3.2 Particle mobility classification by machine learning

We now investigate whether the propensity of a particle to be mobile or immobile (i.e., a dynamic property) could in some ways be encoded in its static initial structure (before loading). To this end, following the example of the softness approach [12,22], we construct by machine learning a structural quantity that is correlated to the propensity of a particle to exhibit a local rearrangement upon creep deformation. Briefly, we first construct a dataset composed of ~17,000 particles obtained from 10 independent creep simulations (with 10 distinct initial configurations). Each system exhibits a similar distribution of D^2_{min}/σ^2 at the end of creep simulation. From this dataset, 7 configurations serve as training set, while the remaining 3 configurations are used as test set. Each particle is classified as mobile ($D^2_{min}/\sigma^2 \geq D^2_{min,0}/\sigma^2$) or immobile ($D^2_{min}/\sigma^2 < D^2_{min,0}/\sigma^2$) based on its final normalized non-affine squared displacement (at the end of the creep simulation). We then calculate a series of structural features for each particle based on the initial static structure

(before any stress is applied). In detail, we calculate for each particle i a series of N_r radial order parameters $G(i; r)$ associated with different distances r:

$$G(i;r) = \sum_j e^{-(R_{ij}-r)^2/L^2} \quad \textbf{Eq. (11-5)}$$

where j refers to the neighbor particles of i within a cutoff distance R_G (here, $R_G = 6\sigma$ [22]), R_{ij} is the distance between the particles i and j, and L is the standard deviation of the Gaussian functions centered around r (here, $L = 0.04\sigma$ [22]). In short, this metric is related to the local density of neighbors at a distance r from the central particle i, as averaged over a shell with a typical thickness L. We calculate for each particle i these N_r order parameters for varying r distances (ranging from 0.6σ to 3σ with an increment of 0.04σ [22]). All these features are standardized prior to any training [46]. Altogether, the ensemble of these metrics offers an unbiased fingerprint of the local radial order around each individual particle. We then train a classifier to identify the optimal hyperplane separating mobile from immobile particles within the N_r-dimensional space associated with the values of the N_r radial order parameters.

In contrast to the original softness approach that uses both radial and angular order parameters as input features [12,22], we solely focus on features capturing radial 2-body correlations around each particle. This is key to ensure that the new softness metric remains highly interpretable (see Sec. 11.4.1). Note that, since the particles are monodisperse and do not exhibit any bond directionality, the incorporation of angular 3-body order parameters does not notably increase the accuracy of the classification model. In that regard, limiting the number of input features also allows us to ensure that the model does not become overfitted. Moreover, unlike the original softness approach based on the Support Vector Machine (SVM) classifier [12,22], we adopt logistic regression to build the classifier [46]. This classifying technique offers great model simplicity, accuracy, and interpretability. Indeed, logistic regression directly provides the

probability of a given particle to be mobile or immobile. In addition, it embeds regularization to limit the risk of overfitting. Importantly, the classification hyperplane determined by logistic regression is linear, which makes it possible to easily assess the importance of each feature. We also expect that the linear nature of the hyperplane is key to enhance the extrapolability of the classification model.

Figure 11-2a illustrates the outcome of the classification, where we select as horizonal and vertical axis the two most influential features of the classification model (see Sec. 11.4.1) in order to illustrate a two-dimensional projection of the positions of the particles in the N_r-dimensional feature space. Each particle is then colored based on its relative non-affine squared displacement. Finally, Fig. 11-2a shows the hyperplane identified by logistic regression, which effectively discriminates mobile from immobile particles. Notably, we find that, based on the knowledge of the structural features, this classifier properly classifies particles as mobile or immobile with an accuracy of 75% and 70% for the training and test sets, respectively. Interestingly, this signals that the propensity for particles to dynamically rearrange during the long-time creep of the gel is largely encoded in its initial static structure (before any stress is applied).

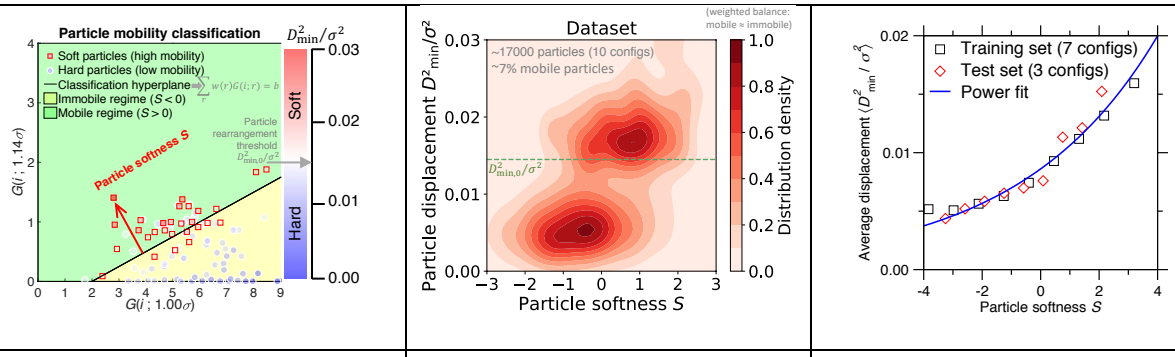
11.3.3 Machine-learned structural metric governing particles' dynamics

The softness S of each particle is then defined as the orthogonal distance from the hyperplane to its position in the N_r-dimensional feature space, wherein mobile (soft) and immobile (hard) particles are associated with positive and negative values of S, respectively. Figure 11-2b shows the distribution density of the particles' normalized non-affine squared displacement $D^2_{\text{min}}/\sigma^2$ (at the end of the creep simulation) and their initial softness S (at the beginning of the simulation, before any stress is applied). We find that, based on the softness sign ($S > 0$ or $S < 0$),

the mobile particles ($D^2_{min}/\sigma^2 \geq D^2_{min,0}/\sigma^2$) can be well discriminated from the immobile particle ($D^2_{min}/\sigma^2 < D^2_{min,0}/\sigma^2$). Further, Figure 11-2c shows the final average normalized non-affine squared displacement $\langle D^2_{min}/\sigma^2 \rangle$ of the particles as a function of their softness S, both for the training and test sets. We find that the normalized non-affine squared displacement of the particles features a power-law dependence on softness. Namely, in addition of properly discriminating mobile from immobile particles, the softness metric also offers some information on the magnitude of the displacement—that is, the particles that exhibit the largest reorganization upon creep are associated with the largest softness values, and vice versa. This power-law correlation is likely to be rooted in the fact the particle dynamics is encoded in the topography of energy landscape of the initial static gel structure (see Sec. 11.4.3).

Notably, the degree of correlation between softness and particle dynamics during creep remains high for particles belonging to the test set. Figure 11-3a offers a snapshot of the predicted softness of an initial static gel configuration in the test set. The distribution of softness (both for all particles and for mobile particles in the gel) is provided in Fig. 11-3b. We find that the classification accuracy is satisfactory as ~76% of the mobile particles indeed exhibit a positive softness ($S > 0$). We then calculate the probability of a particle to rearrange $P_R(S)$ as a function of its initial softness S in the gel (see Fig. 11-3c). Interestingly, we find that $P_R(S)$ exhibits an exponential dependance on the softness metric S, wherein the larger the softness is, the more likely the particle is to rearrange. This relationship can be formulated as an Arrhenius-like behavior in Eq. (11-6) [47]:

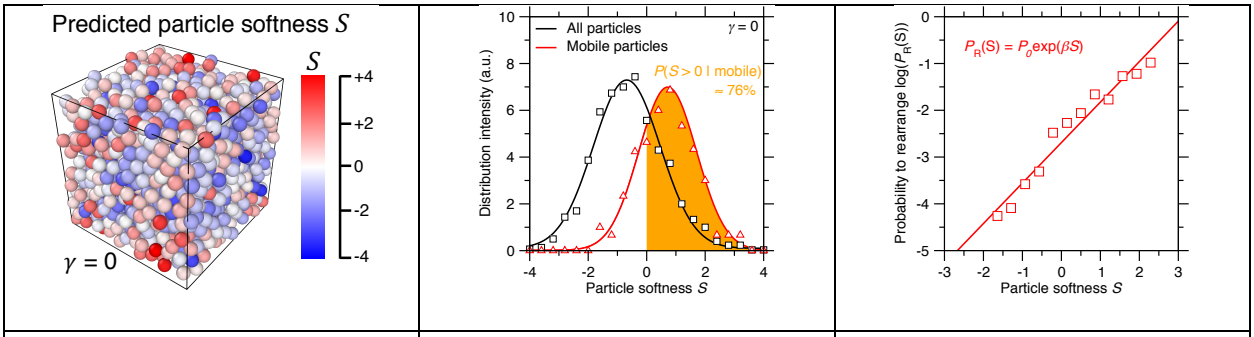
$$P_R(S) = P_0 \exp(\beta S) \qquad \textbf{Eq. (11-6)}$$



(a) **(b)** **(c)**

Figure 11-2: (a) Illustration of the classifier model, wherein the position of each particle is determined from the values of the two most influential structural features used for the classification, i.e., the order parameters $G(i; r)$ calculated at $r_0 = 1.00\sigma$ and $r_1 = 1.14\sigma$. The color of each particle denotes its relative non-affine squared displacement (D^2_{min}/σ^2). The black line represents the projection of the hyperplane identified by logistic regression in this 2-dimensional space. **(b)** Distribution density of the particles' normalized non-affine square displacement (D^2_{min}/σ^2) (at the end of the creep simulation) and initial softness (S), wherein the softness of each particle is defined as the orthogonal distance from the hyperplane to its position in the N_r-dimensional feature space (see panel a). The dataset consists of 10 creep simulations (~1700 particles and ~7% mobile particles per configuration), wherein 7 final configurations serve as training set and the rest 3 configurations are test set. The green dash line indicates the threshold ($D^2_{min,0}/\sigma^2$) of particle rearrangement. For illustration purposes, the density of mobile particles is rescaled to ensure balance with the number of immobile particles. **(c)** Final average normalized non-affine squared displacement $\langle D^2_{min}/\sigma^2 \rangle$ of the particles of the training and test sets (at the end of the creep simulation) as a function of their initial softness. The blue line is a power fit to guide the eye.

where P_0 and β are some fitting parameters. This exponential relationship between $P_R(S)$ and S suggests that the structural quantity S is closely related to (and might be indicative of) the energy barrier associated with particle rearrangements [23,26] (see Sec. 11.4.4). Note that softness is calculated based on the sole knowledge of the initial structure, whereas the normalized non-affine squared displacement is computed at the end of the simulation (i.e., after the gel has creeped to exhibit a shear strain of about 1%). The high degree of correlation between initial softness and final normalized non-affine squared displacement clearly illustrates the intimate link between the initial static structure of the gel and its long-time creep dynamics.



(a) **(b)** **(c)**

Figure 11-3: (a) Snapshot of the predicted particles' softness of an initial static gel (shear strain $\gamma = 0$) in the test set. **(b)** Distribution of the softness of all particles (black) and mobile particles (red) in the gel. The orange area represents the fraction of properly predicted soft particles ($S > 0$) within the mobile particles. **(c)** Logarithm of the probability of a particle to rearrange upon creep ($D^2_{min}/\sigma^2 \geq D^2_{min,0}/\sigma^2$) $\log(P_R(S))$ as a function of its initial softness S in the initial gel structure. The red line is an exponential fit following Eq. (11-6).

11.4 Discussion

11.4.1 Structural interpretation of "particle softness"

We now discuss the structural interpretation of the machine-learned softness metric. As a key advantage of our approach, using the radial order parameters $G(i; r)$ as sole features of the classifier and adopting logistic regression make the softness metric that is constructed herein highly interpretable. Indeed, the hyperplane created by logistic regression can be expressed as a linear equation of each of the features as (see Fig. 11-2a):

$$\mathbf{\Sigma}_r w(r) G(i; r) = b \qquad \textbf{Eq. (11-7)}$$

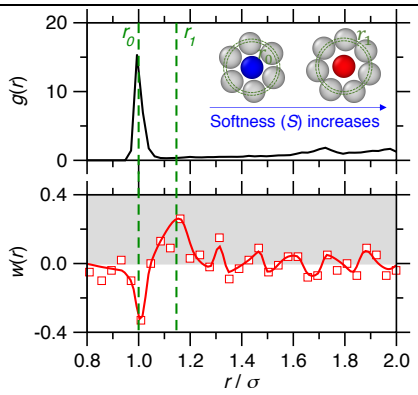
wherein $w(r)$ and b are the coefficients and the bias of the logistic regression model, respectively. Since the features $G(i; r)$ are standardized, the coefficients are directly indicative of the relative importance of each feature in the classification. Namely, a large absolute value for $w(r)$ denotes that the hyperplane is fairly orthogonal to the axis associated with the corresponding feature $G(i; r)$. In addition, the sign of the coefficients is informative, since positive and negative values for $w(r)$ indicate that increasing values of the feature $G(i; r)$ tend to result in increased and decreased softness values, respectively.

Figure 11-4a shows the coefficients $w(r)$ of the logistic regression classifier as a function of the distance r, wherein the absolute value of $w(r)$ denotes how influential the feature $G(i; r)$ is. We find that the two most influential features are associated with the distances $r_0 = \sigma$ and $r_1 = 1.14\sigma$. A visual inspection of the pair distribution function of the gel (see the upper panel of Fig. 11-4a) reveals that r_0 represents the average interparticle bond distance (i.e., equilibrium position of particle interaction energy), while r_1 represents a distance between the 1st and the 2nd coordination shells. This indicates that the local density of neighbors centered around these distances plays a critical role in discriminating mobile from immobile particles. On the one hand, the local density of atoms at $r = r_0$ is related to the coordination number (CN) of the central particle when the neighbors are in contact with the central particle (see blue particle in Fig. 11-4a). The

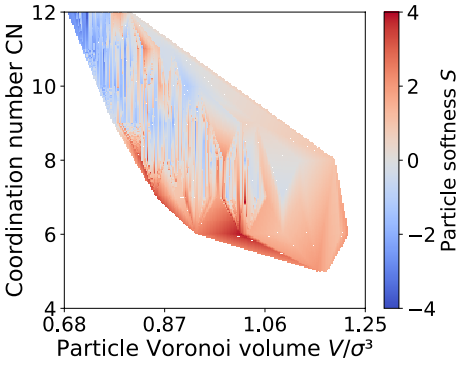
fact that $w(r_0) < 0$ indicates that large CNs tend to result in lower softness values. On the other hand, we interpret the distance r_1 as that wherein neighbors are located when there is coordination mismatch around the central particle (e.g., an excess of neighbors, see red particle in Fig. 11-4a). The fact that $w(r_1) > 0$ indicates that the presence of such coordination mismatch tends to result in higher softness values. These results are consistent with free volume theory (FVT) [8,9]. Indeed, closed-packed structures with a large number of atoms at $r = r_0$ are associated with low local free volume, wherein the atoms exhibit very limited mobility. In contrast, more disordered structures exhibiting notable coordination mismatch tend to show larger local free volume (see red particle in Fig. 11-4a), which facilitates particle mobility.

The absolute values of the coefficients $w(r)$ associated with other distances are notably lower and, hence, the local density of neighbors at such distance has a smaller influence on the outcome of the classification. The features associated with these other distances are nevertheless important for the accuracy of the classifier. Indeed, we find that, even though our results suggest that the local free volume plays an important role in creep dynamics, the classification model trained based on the sole knowledge of the particle Voronoi volume offers a limited accuracy of ~60% as compared to that offered by the softness metric (~75% accuracy). Figure 11-4b illustrates the dependence of softness on the particles' coordination number CN and Voronoi volume V. Overall, larger CN and smaller V values tend to favor smaller softness. However, we nevertheless observe that softness is a complex, nonmonotonic function of CN and V. We note that training a classifier based on the sole knowledge of the particles' CN only yields an accuracy of ~55%, wherein both low and high-coordination particles are very likely to be classified as soft particles, so that the soft vs. hard nature of particles cannot simply be inferred based on Maxwell criterion on stability [48,49]. Similarly, only using $G(i; r_0)$ and $G(i; r_1)$ as input features yields a very limited

accuracy of ~50%. This exemplifies the benefit of using an unbiased machine learning approach to build the set of input features, since intuitive structural features show only limited correlation with dynamical properties.



(a)



(b)

Figure 11-4: (a) Weight coefficient $w(r)$ of classification hyperplane (see Fig. 11-2(a)) at different normalized distances r/σ. The red line is to guide the eyes. The pair distribution function $g(r)$ of the gel is added in the top panel as reference. The distances $r_0 = 1.00\sigma$ and $r_1 = 1.14\sigma$ are associated with the most influential input features of the classifier, i.e., the $w(r)$ coefficients showing maximum absolute value. The inset illustrates the local environments around (i) "hard" particles (blue), wherein the neighbors are located at $r_0 = 1.00\sigma$ and (ii) "soft" particles (red), wherein the neighbors are located at $r_1 = 1.14\sigma$. **(b)** Particle softness S as a function of their normalized particle Voronoi volume V/σ^3 and coordination number CN. The color coding is based on a linear interpolation between the datapoints in the particle dataset.

11.4.2 Linking particle dynamics to macroscopic deformation

We now look into the nature of the relationship between particle-level softness (see Fig. 11-3c) and the macroscopic creep of the gel (see Fig. 11-1a). To this end, we consider the average

softness ‹*S*› of the system—i.e., as averaged over all the particles—and assess how this quantity is evolving as the gel gradually undergoes creep (see the inset of Fig. 11-5). We find that ‹*S*› exhibits a logarithmic decay upon creep, which echoes the logarithmic increase of the macroscopic strain of the gel upon creep (see Fig. 11-1a). In fact, as shown in Fig. 11-5, we observe that the macroscopic creep rate $\dot{\gamma}$ of the gel exhibits an exponential dependence on ‹*S*› as:

$$\dot{\gamma} = \dot{\gamma}_0 \exp(\alpha \langle S \rangle) \qquad \textbf{Eq. (11-8)}$$

where $\dot{\gamma}_0$ and α are some fitting parameters. Note that this exponential relationship is not affected by the system size. This indicates that the dynamics of creep at the macroscopic scale is closely related to the variation in softness at the particle level. This can be understood as follows. The gradual decay of softness indicates that, upon creep, particles reorganize from "soft" (i.e., high *S*) to "harder" (i.e., lower *S*) local environments (see the schematics in Fig. 11-4a). In turn, as the softness of a particle decreases, so does its propensity to exhibit any further reorganization. This process explains why the creep rate gradually slows down—since the particles gradually become harder and harder and, hence, less prone to reorganizations. It is worthwhile to point out that, although our softness results illustrate a strong correlation between the initial static structure and the early-stage creep dynamics, it remains unclear whether the softness approach presented herein could describe longer-term effects (e.g., final avalanche) based on the static initial structure since the system tends to lose the memory of its initial configuration after experiencing significant deformations [11,37,38,50]. Here, the exponential dependence of the creep rate on the average softness is a macroscopic manifestation of the particle-level exponential dependence of particle dynamics on its softness in Eq. (11-6) (see Fig. 11-3c), which suggests that ‹*S*› (or, more accurately, the opposite thereof) captures an effective average energy barrier for creep [2,8]—wherein low ‹*S*›

values (i.e., “hard” structures) are indicative of high energy barriers for particle rearrangements, and vice versa (see below) [23,26].

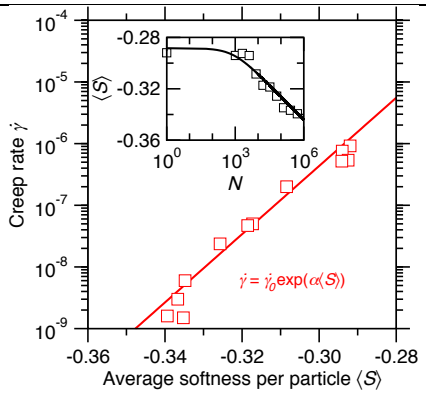


Figure 11-5: Macroscopic creep rate $\dot{\gamma}$ of the gel as a function of the average softness $\langle S \rangle$ of the particles. The red line is an exponential fit following Eq. (11-8). The inset shows the evolution of $\langle S \rangle$ in the gel upon creep.

11.4.3 The energy landscape governs the particle dynamics during creep

Since the results shown in Fig. 11-5 suggest that softness may be capturing the effective energy barrier that is accessible to the particles during creep, we now further investigate the linkage between creep dynamics and potential energy landscape (PEL) topography [47]. Figure 11-6a offers a schematic of the local PEL that is accessible to an initial static gel (before any stress is applied). The initial configuration is located at a local minimum of the PEL. Starting from this initial position, the ARTn algorithm searches for saddle points around the local minimum, which are associated with physically-meaningful rearrangements for a target particle [41]. This allows us to compute the distribution of the energy barriers that are locally accessible to each particle (i.e., the energy difference between the identified saddle point and initial minimum, see Sec. 11.2.5). Based on this analysis, we then calculate, for each particle, the average value E_{ave} of the energy barriers that are accessible to this particle.

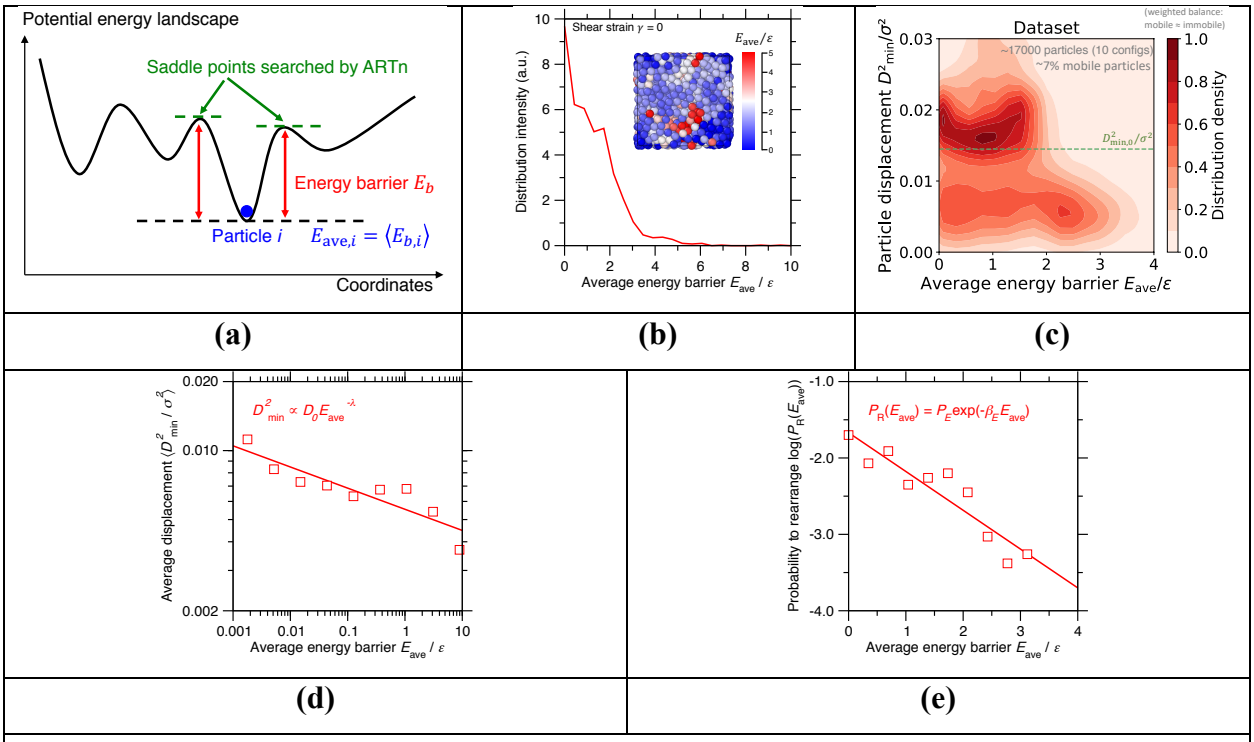


Figure 11-6: (a) Schematic illustrating the local potential energy landscape (PEL) of an initial static gel. The gel is initially located at a local minimum of the PEL. Starting from this minimum position, the activation-relaxation nouveau (ARTn) algorithm searches the saddle points that are locally accessible to target particles [41] (see Methods section for details). E_{ave} is the average value of the energy barriers that are accessible to a given particle. **(b)** Distribution of the normalized average energy barrier E_{ave}/ε of the particles in the initial static gel (before any stress is applied). The inset shows the associated gel configuration, wherein the color of the particles denotes E_{ave}/ε. **(c)** Distribution density of the particles' normalized non-affine square displacement (D^2_{min}/σ^2) (at the end of the creep simulation) and initial normalized average energy barrier (E_{ave}/ε). **(d)** Final average normalized non-affine squared displacement $\langle D^2_{min}/\sigma^2 \rangle$ of the particles in the gel (at the end of the creep simulation) as a function of their initial normalized average energy barrier (E_{ave}/ε). The red line is a power-law fit. **(e)** Logarithm of the probability of a particle to rearrange upon creep ($D^2_{min}/\sigma^2 \geq D^2_{min,0}/\sigma^2$)

$\log(P_R(E_{ave}))$ as a function of its initial normalized average energy barrier E_{ave}/ε in the gel. The red line is an exponential fit following Eq. (11-9).

Figure 11-6b shows the distribution of the normalized average energy barrier E_{ave}/ε of the particles in the initial static gel. We find that the distribution decreases with increasing energy barrier value and exhibits a long tail toward high energy barriers. Figure 11-6c shows the distribution density of the particles' normalized non-affine squared displacement D^2_{min}/σ^2 (at the end of the creep simulation) and initial normalized average energy barrier E_{ave}/ε. We observe the existence of an anticorrelation between displacement and average energy barrier—which is a natural consequence of the fact that particles that are surrounded only by large energy barriers (i.e., rough local energy landscape) are trapped around their local minimum and unable to reorganize [47]. However, for low E_{ave}/ε values, we find that only a small fraction of the particles tends to exhibit a large displacement (i.e., to be mobile). This can be explained based on the spatial heterogeneity of the D^2_{min}/σ^2 and E_{ave}/ε fields in the gel (see discussion on the point in Sec. 11.4.4).

Figure 11-6d shows the final average normalized non-affine squared displacement $\langle D^2_{min}/\sigma^2 \rangle$ of the particles in the gel (at the end of the creep simulation) as a function of their normalized average energy barrier E_{ave}/ε (in the initial configuration, before any stress is applied). Interestingly, we find the existence of a power-law relationship between $\langle D^2_{min}/\sigma^2 \rangle$ and E_{ave}/ε, wherein larger E_{ave}/ε tend to result in smaller D^2_{min}/σ^2 values, and vice versa (see Fig. 11-6d). This result echoes the power-law relationship between $\langle D^2_{min}/\sigma^2 \rangle$ and softness S previously highlighted in Fig. 11-2c. The harmony between these trends suggests the existence of a potential causal relationship between softness S and average energy barrier E_{ave} (see Sec. 11.4.4). In addition, these results also echo findings from a recent study, which reported the existence of a power-law

relationship between particle dynamics and energy barrier in metallic glasses [42,43]. This suggests that this power-law relationship between particle dynamics and energy barrier (or particle softness) might be a generic feature of disordered systems, independently of whether the particle reorganizations are caused by creep or not.

Finally, we evaluate the probability of a particle to rearrange $P_R(E_{\text{ave}})$ as a function of its initial average energy barrier E_{ave} (see Fig. 11-6e). We find $P_R(E_{\text{ave}})$ follows an Arrhenius-like exponential dependence on E_{ave} as [47]:

$$P_R(E_{\text{ave}}) = P_E \exp(-\beta_E E_{\text{ave}}) \qquad \textbf{Eq. (11-9)}$$

where P_E and β_E are some fitting parameters. Namely, the larger the average energy barrier is, the less likely the particle is to reorganize. Overall, these results indicate that the dynamics of a gel upon creep is largely encoded in the topography of its initial energy landscape, before any load is applied. The harmony between Eqs. (11-6) and (11-9)—which both exhibit an Arrhenius form—suggests a strong correlation between the softness S and average energy barrier E_{ave} fields (see below).

11.4.4 Mapping "particle softness" to energy barrier

Finally, we interrogate the existence of a causal correlation between softness S and average energy barrier E_{ave}. Figure 11-7a illustrates the spatial correlation between the fields of interest herein: (i) the final non-affine squared displacement (D^2_{min}), (ii) the initial softness (S), and (iii) the initial average energy barrier (E_{ave}). We find that the three fields—i.e., the dynamics field (D^2_{min}), the structural field (S), and the local potential energy landscape field (E_{ave})—all show a strong degree of spatial heterogeneity, but are fairly correlated to each other. Namely, the regions associated with low mobility tend to match with those presenting high energy barriers and low softness, and vice versa.

Figure 11-7b shows the distribution density of the particles' initial normalized average energy barrier (E_{ave}/ε) and initial softness (S). Overall, we observe that the particles associated high softness values (i.e., "soft" mobile particles, $S > 0$) tend to exhibit fairly low average energy barriers. However, we note that the correlation between softness and average energy barriers is not as strong as that observed between softness and displacement, or energy barriers and displacement (see Figs. 11-2b and 11-6c). Nevertheless, a stronger correlation between softness and energy landscape emerges when averaging these fields over groups of particles featuring fairly similar softness. To this end, Fig. 11-7c shows the initial average normalized energy barrier $\langle E_{ave}/\varepsilon \rangle$ of the particles in the gel (before any stress is applied) as a function of their initial softness S. Interestingly, we find that $\langle E_{ave}/\varepsilon \rangle$ is linearly related to the opposite of softness ($-S$, or "hardness" H) as:

$$E_{\mathrm{ave}}(H) = kH + b \qquad \textbf{Eq. (11-10)}$$

where k and b are some fitting parameters. This shows that the topography of the energy landscape is largely encoded in the structure of the gel. Importantly, these results demonstrate that, when averaged over groups of particles, the average softness indeed offers a purely structural metric that successfully captures the average height of the energy barriers that are locally accessible, which, in turn, controls the particles' propensity to reorganize upon creep [23,26]. It is notable that the softness metric is able to successfully capture a structural fingerprint for the topography of the potential energy landscape since this machine-learned quantity is not trained for that purpose (that is, the model is never exposed to the energy barriers during its training). As such, the correlation between softness and average energy barrier offers an independent validation of the soundness of this approach—and suggests that the softness metric that is extracted herein indeed shows a "real" physical meaning.

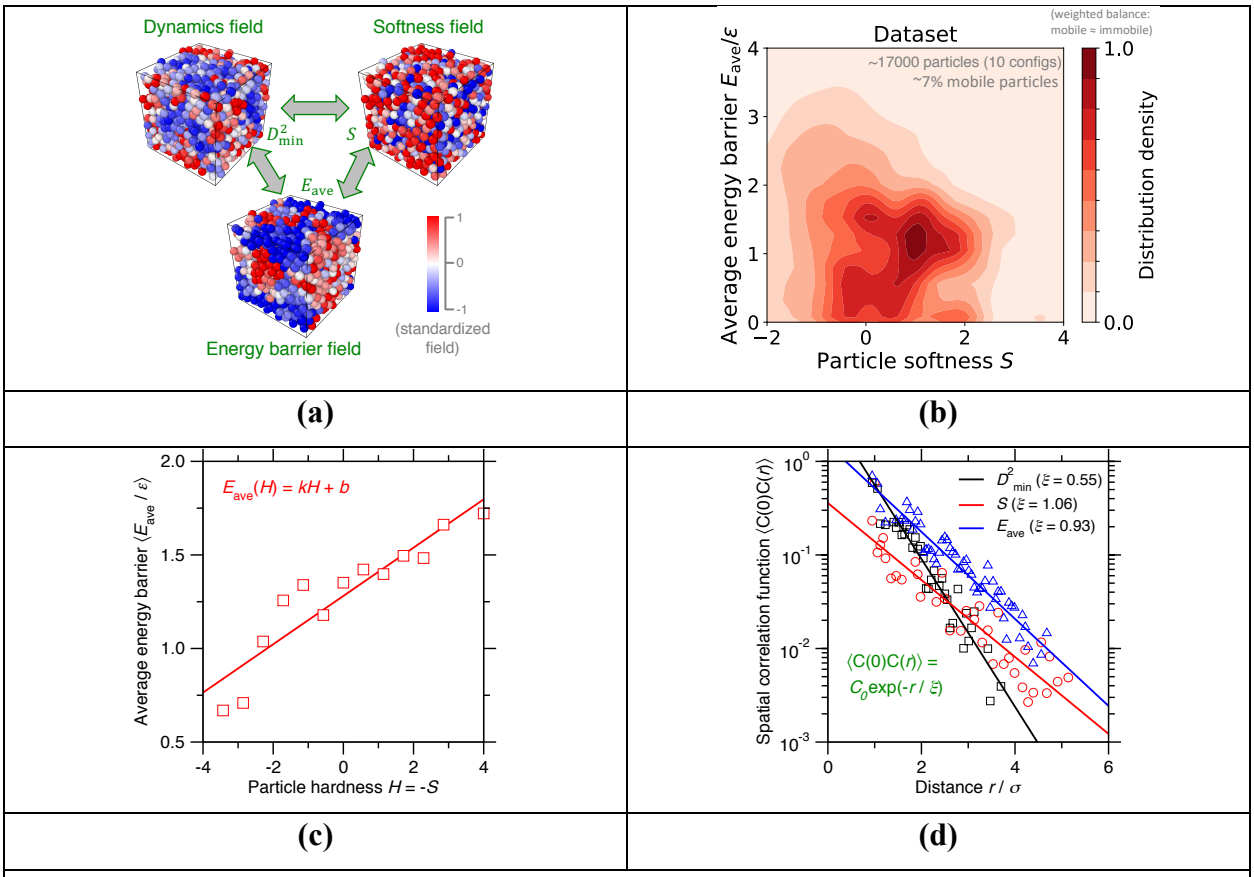


Figure 11-7: (a) Illustration of the spatial correlation between the fields of the final non-affine squared displacement (D^2_{min}), the initial softness (S), and the initial average energy barrier (E_{ave}). The particles are colored based on their standardized value in the corresponding field. **(b)** Distribution density of the particles' initial normalized average energy barrier (E_{ave}/ε) and initial softness (S). **(c)** Initial average normalized energy barrier ($\langle E_{ave}/\varepsilon \rangle$) of the particles in the gel (before any stress is applied) as a function of their initial hardness ($H = -S$). The red line is a linear fit following Eq. (11-10). **(d)** Spatial correlation function $\langle C(0)C(r) \rangle$ of the displacement field (D^2_{min}, black), the softness field (S, red), and the energy barrier field (E_{ave}, blue) in the gel. Note that the field value C (i.e., D^2_{min}, S, E_{ave}) is standardized for the calculation. The lines are exponential fits following $\exp(-r/\xi)$, where ξ is the characteristic correlation length.

Note that, in this analogy between softness and energy barriers, the energy barriers that are captured by softness are not overcome by thermal effects, but by the applied stress—which provides an elastic energy that enables particles to jump over these barriers [2,8]. An interesting atomic picture behind the link between energy landscape topography, applied external stress (or strain), and resulting particle hoping is offered by the trap model from Barrat *et al.*—which describes the energy landscape as a landscape of "traps," wherein an external stress can facilitate particle hopping from trap to trap by deforming the local landscape [47]. In that regard, our results suggest that softness might serve as a proxy for the average height of the energy barriers that separate the traps within the energy landscape.

To further explore the degree of spatial heterogeneity in the three fields considered herein, we compute the spatial correlation function ‹C(0)C(*r*)› for each field (see Fig. 11-7d), where C(*r*) is the normalized fluctuation in the field, i.e., the standardized field value (D^2_{min}, S, and E_{ave}) of a particle at distance *r* from a central particle. The spatial correlation function ‹C(0)C(r)› is computed by averaging over all particles separated by a distance *r*. We then infer the characteristic correlation length ξ associated with each field by fitting the spatial correlation function as [24]:

$$< \mathrm{C}(0)\mathrm{C}(r) > = C_0 \exp\left(-r/\xi\right) \qquad \textbf{Eq. (11-11)}$$

where C_0 is a fitting coefficient. We find that both the softness field and the energy barrier field show a similar correlation length ξ that is close to 1 (i.e., the typical radius of the first coordination shell). This harmony further supports the close relationship between softness and energy barriers. In contrast, the dynamics field shows a correlation length of $\xi \approx 0.5$ (i.e., the typical radius of a particle). This indicates that the typical lengthscale associated with particle displacements is notably lower as compared to that associated with the softness/energy barriers fields. The fact that the lenthscale associated with displacements is lower than that of the energy barrier field likely

explains why only a small fraction of the particles showing low E_{ave}/ε values also feature large displacements (see Fig. 11-6c). This partial decorrelation between the spatial distributions of displacement and energy barriers (or softness) may be a consequence of the fact that, here, displacement is induced by stress rather than being fully spontaneous. Consequently, "soft" particles that have access to low energy barriers may nevertheless not exhibit any notable displacement if the direction of the imposed stress does not match with any of the accessible low-energy saddle point pathways.

11.5 Conclusions

Overall, these results highlight the close correlation between (i) static structure (as captured by softness), (ii) static potential energy landscape topography (as captured by the average height of the energy barriers that are accessible to the particles), (iii) particle dynamics (as captured by the non-affine squared displacement), and (iv) macroscopic deformation (as captured by the creep rate). It is notable that our approach allows us to predict the long-time dynamics of the particles upon long-term creep deformations while solely relying on the knowledge of the initial static structure before any stress is applied. The accessible interpretation of the softness metric defined herein (see Fig. 11-4a) suggests that the degree of structural disorder—and especially the existence of coordination mismatches—plays a key role in governing the creep dynamics of gels. This indicates that order-disorder engineering of gel structures offers a potential path to develop new gel formulations with tailored creep response under sustained load.

11.6 References

[1] V.B. Nguyen, T. Darnige, A. Bruand, E. Clement, Creep and Fluidity of a Real Granular Packing near Jamming, Physical Review Letters. 107 (2011). https://doi.org/10.1103/PhysRevLett.107.138303.

[2] M. Bauchy, M. Wang, Y. Yu, B. Wang, N.M.A. Krishnan, E. Masoero, F.-J. Ulm, R. Pellenq, Topological Control on the Structural Relaxation of Atomic Networks under Stress, Physical Review Letters. 119 (2017). https://doi.org/10.1103/PhysRevLett.119.035502.

[3] M. Siebenbürger, M. Ballauff, Th. Voigtmann, Creep in Colloidal Glasses, Physical Review Letters. 108 (2012). https://doi.org/10.1103/PhysRevLett.108.255701.

[4] J.-P. Poirier, Creep of Crystals: High-Temperature Deformation Processes in Metals, Ceramics and Minerals, Cambridge University Press, 1985.

[5] W.N. Findley, F.A. Davis, Creep and Relaxation of Nonlinear Viscoelastic Materials, Courier Corporation, 2013.

[6] Y.M. Joshi, Dynamics of Colloidal Glasses and Gels, Annu. Rev. Chem. Biomol. Eng. 5 (2014) 181–202. https://doi.org/10.1146/annurev-chembioeng-060713-040230.

[7] M. Vandamme, F.-J. Ulm, Nanoindentation investigation of creep properties of calcium silicate hydrates, Cement and Concrete Research. 52 (2013) 38–52. https://doi.org/10.1016/j.cemconres.2013.05.006.

[8] H. Liu, S. Dong, N.M.A. Krishnan, E. Masoero, G. Sant, M. Bauchy, Long-term creep deformations in colloidal calcium–silicate–hydrate gels by accelerated aging simulations, Journal of Colloid and Interface Science. 542 (2019) 339–346. https://doi.org/10.1016/j.jcis.2019.02.022.

[9] M. Vandamme, F.-J. Ulm, Nanogranular origin of concrete creep, PNAS. 106 (2009) 10552–10557. https://doi.org/10.1073/pnas.0901033106.

[10] K. Ioannidou, K.J. Krakowiak, M. Bauchy, C.G. Hoover, E. Masoero, S. Yip, F.-J. Ulm, P. Levitz, R.J.-M. Pellenq, E.D. Gado, Mesoscale texture of cement hydrates, PNAS. 113 (2016) 2029–2034. https://doi.org/10.1073/pnas.1520487113.

[11] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E.D. Cubuk, S.S. Schoenholz, A. Obika, A.W.R. Nelson, T. Back, D. Hassabis, P. Kohli, Unveiling the predictive power of static structure in glassy systems, Nat. Phys. 16 (2020) 448–454. https://doi.org/10.1038/s41567-020-0842-8.

[12] E.D. Cubuk, R.J.S. Ivancic, S.S. Schoenholz, D.J. Strickland, A. Basu, Z.S. Davidson, J. Fontaine, J.L. Hor, Y.-R. Huang, Y. Jiang, N.C. Keim, K.D. Koshigan, J.A. Lefever, T. Liu, X.-G. Ma, D.J. Magagnosc, E. Morrow, C.P. Ortiz, J.M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K.N. Nordstrom, P.E. Arratia, R.W. Carpick, D.J. Durian, Z. Fakhraai,

D.J. Jerolmack, D. Lee, J. Li, R. Riggleman, K.T. Turner, A.G. Yodh, D.S. Gianola, A.J. Liu, Structure-property relationships from universal signatures of plasticity in disordered solids, Science. 358 (2017) 1033–1037. https://doi.org/10.1126/science.aai8830.

[13] T. Aste, M. Saadatfar, T.J. Senden, Geometrical structure of disordered sphere packings, Phys. Rev. E. 71 (2005) 061302. https://doi.org/10.1103/PhysRevE.71.061302.

[14] A. Bunde, S. Havlin, Fractals and Disordered Systems, Springer Science & Business Media, 2012.

[15] Q. Wang, A. Jain, A transferable machine-learning framework linking interstice distribution and plastic heterogeneity in metallic glasses, Nat Commun. 10 (2019) 1–11. https://doi.org/10.1038/s41467-019-13511-9.

[16] M.A. Klatt, J. Lovrić, D. Chen, S.C. Kapfer, F.M. Schaller, P.W.A. Schönhöfer, B.S. Gardiner, A.-S. Smith, G.E. Schröder-Turk, S. Torquato, Universal hidden order in amorphous cellular geometries, Nature Communications. 10 (2019) 811. https://doi.org/10.1038/s41467-019-08360-5.

[17] M. Mungan, S. Sastry, K. Dahmen, I. Regev, Networks and Hierarchies: How Amorphous Materials Learn to Remember, Phys. Rev. Lett. 123 (2019) 178002. https://doi.org/10.1103/PhysRevLett.123.178002.

[18] R.L. Jack, A.J. Dunleavy, C.P. Royall, Information-Theoretic Measurements of Coupling between Structure and Dynamics in Glass Formers, Phys. Rev. Lett. 113 (2014) 095703. https://doi.org/10.1103/PhysRevLett.113.095703.

[19] A. Widmer-Cooper, P. Harrowell, H. Fynewever, How Reproducible Are Dynamic Heterogeneities in a Supercooled Liquid?, Phys. Rev. Lett. 93 (2004) 135701. https://doi.org/10.1103/PhysRevLett.93.135701.

[20] G. Biroli, Machine learning glasses, Nat. Phys. 16 (2020) 373–374. https://doi.org/10.1038/s41567-020-0873-1.

[21] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature. 559 (2018) 547. https://doi.org/10.1038/s41586-018-0337-2.

[22] E.D. Cubuk, S.S. Schoenholz, J.M. Rieser, B.D. Malone, J. Rottler, D.J. Durian, E. Kaxiras, A.J. Liu, Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods, Physical Review Letters. 114 (2015). https://doi.org/10.1103/PhysRevLett.114.108001.

[23] S.S. Schoenholz, E.D. Cubuk, D.M. Sussman, E. Kaxiras, A.J. Liu, A structural approach to relaxation in glassy liquids, Nature Physics. 12 (2016) 469–471. https://doi.org/10.1038/nphys3644.

[24] E.D. Cubuk, S.S. Schoenholz, E. Kaxiras, A.J. Liu, Structural Properties of Defects in Glassy Liquids, J. Phys. Chem. B. 120 (2016) 6139–6146. https://doi.org/10.1021/acs.jpcb.6b02144.

[25] D.M. Sussman, S.S. Schoenholz, E.D. Cubuk, A.J. Liu, Disconnecting structure and dynamics in glassy thin films, PNAS. 114 (2017) 10601–10605. https://doi.org/10.1073/pnas.1703927114.

[26] X. Ma, Z.S. Davidson, T. Still, R.J.S. Ivancic, S.S. Schoenholz, A.J. Liu, A.G. Yodh, Heterogeneous Activation, Local Structure, and Softness in Supercooled Colloidal Liquids, Physical Review Letters. 122 (2019). https://doi.org/10.1103/PhysRevLett.122.028001.

[27] E. Masoero, E. Del Gado, R.J.-M. Pellenq, F.-J. Ulm, S. Yip, Nanostructure and Nanomechanics of Cement: Polydisperse Colloidal Packing, Phys. Rev. Lett. 109 (2012) 155503. https://doi.org/10.1103/PhysRevLett.109.155503.

[28] H. Liu, S. Dong, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Effects of polydispersity and disorder on the mechanical properties of hydrated silicate gels, Journal of the Mechanics and Physics of Solids. 122 (2019) 555–565. https://doi.org/10.1016/j.jmps.2018.10.003.

[29] H. Liu, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Structural percolation controls the precipitation kinetics of colloidal calcium–silicate–hydrate gels, J. Phys. D: Appl. Phys. 52 (2019) 315301. https://doi.org/10.1088/1361-6463/ab217b.

[30] E. Masoero, E.D. Gado, R. J.-M. Pellenq, S. Yip, F.-J. Ulm, Nano-scale mechanics of colloidal C–S–H gels, Soft Matter. 10 (2014) 491–499. https://doi.org/10.1039/C3SM51815A.

[31] K. Ioannidou, R. J.-M. Pellenq, E.D. Gado, Controlling local packing and growth in calcium–silicate–hydrate gels, Soft Matter. 10 (2014) 1121–1133. https://doi.org/10.1039/C3SM52232F.

[32] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[33] Y. Yu, M. Wang, D. Zhang, B. Wang, G. Sant, M. Bauchy, Stretched Exponential Relaxation of Glasses at Low Temperature, Physical Review Letters. 115 (2015). https://doi.org/10.1103/PhysRevLett.115.165901.

[34] Y. Yu, M. Wang, M.M. Smedskjaer, J.C. Mauro, G. Sant, M. Bauchy, Thermometer Effect: Origin of the Mixed Alkali Effect in Glass Relaxation, Phys. Rev. Lett. 119 (2017) 095501. https://doi.org/10.1103/PhysRevLett.119.095501.

[35] P. Richard, M. Nicodemi, R. Delannay, P. Ribière, D. Bideau, Slow relaxation and compaction of granular systems, Nature Materials. 4 (2005) 121–128. https://doi.org/10.1038/nmat1300.

[36] D.J. Lacks, M.J. Osborne, Energy Landscape Picture of Overaging and Rejuvenation in a Sheared Glass, Physical Review Letters. 93 (2004). https://doi.org/10.1103/PhysRevLett.93.255501.

[37] V. Grenard, T. Divoux, N. Taberlet, S. Manneville, Timescales in creep and yielding of attractive gels, Soft Matter. 10 (2014) 1555. https://doi.org/10.1039/c3sm52548a.

[38] C. Liu, K. Martens, J.-L. Barrat, Mean-Field Scenario for the Athermal Creep Dynamics of Yield-Stress Fluids, Phys. Rev. Lett. 120 (2018) 028004. https://doi.org/10.1103/PhysRevLett.120.028004.

[39] M.L. Falk, J.S. Langer, Dynamics of viscoplastic deformation in amorphous solids, Physical Review E. 57 (1998) 7192–7205. https://doi.org/10.1103/PhysRevE.57.7192.

[40] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool, Modelling Simul. Mater. Sci. Eng. 18 (2010) 015012. https://doi.org/10.1088/0965-0393/18/1/015012.

[41] N. Mousseau, L.K. Béland, P. Brommer, J.-F. Joly, F. El-Mellouhi, E. Machado-Charry, M.-C. Marinica, P. Pochet, The Activation-Relaxation Technique: ART Nouveau and Kinetic ART, Journal of Atomic, Molecular, and Optical Physics. 2012 (2012) 1–14. https://doi.org/10.1155/2012/925278.

[42] L. Tang, G. Ma, H. Liu, W. Zhou, M. Bauchy, Bulk Metallic Glasses' Response to Oscillatory Stress Is Governed by the Topography of the Energy Landscape, J. Phys. Chem. B. (2020) acs.jpcb.0c08794. https://doi.org/10.1021/acs.jpcb.0c08794.

[43] L. Tang, H. Liu, G. Ma, T. Du, N. Mousseau, W. Zhou, M. Bauchy, The Energy Landscape Governs Ductility in Disordered Materials, Mater. Horiz. (2021). https://doi.org/10.1039/D0MH00980F.

[44] C.E. Maloney, A. Lemaître, Amorphous systems in athermal, quasistatic shear, Physical Review E. 74 (2006). https://doi.org/10.1103/PhysRevE.74.016118.

[45] C.E. Maloney, D.J. Lacks, Energy barrier scalings in driven systems, Phys. Rev. E. 73 (2006) 061106. https://doi.org/10.1103/PhysRevE.73.061106.

[46] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[47] A. Nicolas, E.E. Ferrero, K. Martens, J.-L. Barrat, Deformation and flow of amorphous solids: Insights from elastoplastic models, Rev. Mod. Phys. 90 (2018) 045006. https://doi.org/10.1103/RevModPhys.90.045006.

[48] J.C. Maxwell, L. *On the calculation of the equilibrium and stiffness of frames*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 27 (1864) 294–299. https://doi.org/10.1080/14786446408643668.

[49] C.F. Moukarzel, Isostatic Phase Transition and Instability in Stiff Granular Materials, Phys. Rev. Lett. 81 (1998) 1634–1637. https://doi.org/10.1103/PhysRevLett.81.1634.

[50] B. Keshavarz, T. Divoux, S. Manneville, G.H. McKinley, Nonlinear Viscoelasticity and Generalized Failure Criterion for Polymer Gels, ACS Macro Lett. 6 (2017) 663–667. https://doi.org/10.1021/acsmacrolett.7b00213.

Chapter 12. Summary and Outlook

12.1 Summary of the Thesis

Overall, owing to its low computation cost and high prediction accuracy, materials modeling has become an efficient alternative to traditional experiments to interrogate materials structure-property relationship and discover new materials [1–3]. Figure 12-1 summarizes the different paradigms for materials discovery [4,5], including Edisonian trial-and-error method (see Sec. 1.1.2), high-throughput virtual screening (see Sec. 1.1.3), machine learning using experimental data (see Sec. 1.3.2), and integration of machine learning and simulations (see Sec. 1.4.1). Clearly, materials modeling is revolutionizing materials discovery paradigms through rationalizing the exploration of vast material design space [4,5]. However, the state-of-the-art materials modeling is facing two grand challenges, i.e., (i) the high complexity of physics laws that govern materials properties (see Sec. 1.3.1), and (ii) the low informativity of experimental dataset (see Sec. 1.3.2) [6,7].

In order to address the two grand challenges facing materials modeling, next-generation materials modeling aims to (i) make the physics simple to facilitate physics-driven modeling (see Chapter 2-4) [8–10], and (ii) make the data informative to facilitate data-driven modeling (see Chapter 5) [7]. In this thesis, I highlight the predictive power of integrating data-driven machine learning (ML) and physics-driven computational simulations to unlock a new era for materials discovery and for next-generation materials modeling (see Chapter 6-11) [11–16]. In detail, on the one hand, simulation can generate large amounts of high-fidelity data that can be used to train machine learning models, which, in turn, can be validated by simulations (Chapter 6) [11]. On the other hand, ML can assist in (i) developing empirical forcefields for accurate and computationally-efficient MD simulations (Chapter 7-9) [12–14], and (ii) “separating the wheat from the chaff” in

large amounts of complex simulation data to gain new insights or generate new knowledge of the underlying physics governing materials behaviors (Chapter 10-11) [15,16].

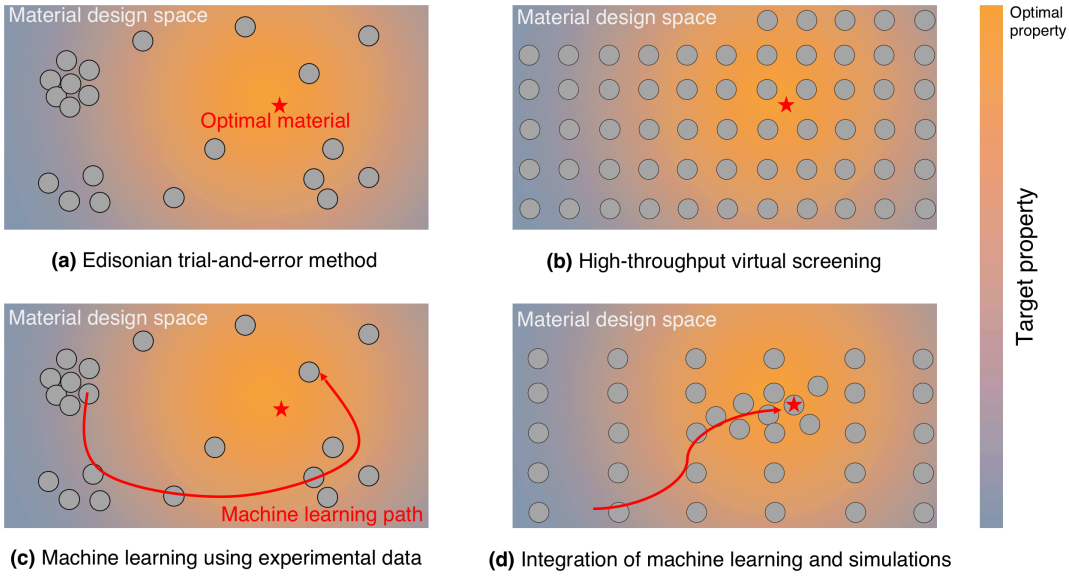


Figure 12-1: Summary of different paradigms for materials discovery, including **(a)** Edisonian trial-and-error method, **(b)** high-throughput virtual screening, **(c)** machine learning using experimental data, and **(d)** integration of machine learning and simulations. The color coding represents the target property in the material design space, and the red star denotes the optimal material exhibiting optimal property. The grey circles are the present datapoints explored by the method, and the red arrow shows the machine learning search path.

12.2 Future Opportunities in Modeling of Disordered Materials

As a future opportunity, I envision that smart closed-loop integrations of ML modeling and simulations will leapfrog materials modeling (see Fig. 12-2). Example of such opportunities are listed in the following:

(i) Machine learning forcefields. Machine-learned forcefields adopt ML regression models to fit the system's potential energy landscape (PEL) as a function of the atom positions,

without the need to rely on any physical or chemical intuition to define a functional format of the empirical forcefield [17,18]. In contrast to empirical forcefields relying on a fixed analytical form, machine-learned forcefields offer a promising pathway to develop new complex forcefields that rely on an ML model to map a given atomic configuration to its potential energy. The promise of machine-learned forcefields is to approach the accuracy of *ab initio* simulations with a computational burden that is more comparable to (although typically higher than) that of classical empirical forcefields [17,18].

(ii) Deciphering complex simulation data by machine learning. Atomic trajectories generated by MD simulations contains all the structural information that govern glassy materials' properties [19]. However, due to the complex, disordered nature of glassy structures [20], it is challenging to "separate the wheat from the chaff," that is, to pinpoint the key structural features that govern materials properties [21]. In that regard, owing to its ability to discover hidden pattern in complex, multi-dimensional data [22], machine learning (ML)—e.g., the recently developed softness approach based on classification-based ML [23]—offers a new opportunity to identify relevant structural patterns in simulated glassy structures [24].

(iii) Combining simulations and machine learning for materials' inverse design. Owing to its economical nature as compared to systematic experiments, high-throughput virtual screening (HTVS, that is, the systematic simulation of a large number of materials) offers an efficient route to identify *in silico* an optimal material featuring an optimal characteristic within a given design space [25]. However, although simulations excel at predicting the properties of a given material as a function of its structure (i.e., "forward prediction"), their application to "inverse design" problems (that is, given an optimal property target, find the best material structure) remains limited by their high computational cost—which prevents the systematic exploration of large design

spaces [11]. To address this issue, machine learning (ML) offers an ideal companion to simulations—since an ML model can learn from a series of simulations and, based on this, recommend what should be next material structure to simulate [26,27]. Such closed-loop integrations of simulation and ML modeling could greatly accelerate the discovery of novel materials featuring desirable properties or functionalities [1–3].

(iv) Leveraging differentiable programming platforms. When integrating simulations and machine learning (ML) models within unified pipelines, different programming languages can present a communication barrier between ML and simulation packages, which often rely on Python and C++/Fortran, respectively [28,29]. In addition, most simulation packages are still rooted in fairly ancient computing paradigms (e.g., with no automated differentiation), which is reminiscent of the state of machine learning before automatic hardware acceleration and differentiation became popular [30–32]. To overcome these frictions, automatic differentiable (auto-diff) programming platforms (e.g., Python JAX [33]) have been recently developed to seamlessly integrate ML and simulations within unified pipelines [11,31]. In contrast to traditional programming platforms that rely on handwritten derivatives (e.g., C++), auto-diff platforms excel at computing on-the-fly the backward gradient of any quantities with no additional computation burden associated with differentiation—an operation that comes with a notable computing time in traditional simulators (e.g., force calculation in MD simulations) [34]. Moreover, simulations built on auto-diff platforms gain backward differentiability, which makes it possible to use their outcomes to directly train machine learning models using for gradient backpropagation. This create new opportunities to train a ML model directly based on differentiable physical knowledge rather than on data [31]. Finally, the auto-diff platforms generally enable native "just-in-time" compilation on high-performance dedicated hardware accelerators, such as graphics processing units (GPU) and tensor

processing units (TPU) [33,35]. Specifically, TPUs are specifically designed as matrix processors and, thanks to their tailored architecture, offer unparalleled performances in deep learning problems (up to 200X faster than GPUs) [35–38]. This could greatly accelerate MD simulations relying on artificial neural networks potentials.

(v) Replacing slow simulations by faster surrogate machine learning simulation engines. Although the development of auto-diff platforms enables differentiable simulations and native hardware acceleration, the computational efficiency of numerical simulations is still limited by the intrinsic computation cost associated with the underlying numerical algorithms (e.g., Newton's law of motion in MD simulations) [39]. The numerical algorithms behind scientific numerical simulations are likely to remain their bottleneck. To mitigate this issue, surrogate machine learning (ML) simulation engines offer a unique, largely untapped opportunity to replace slow simulations so as to accelerate their execution without compromising accuracy [40,41]. Surrogate ML engines can be implemented as artificial neural network (ANN) models, such as convolutional neural network (CNN) [41] or graph neural network (GNN) [42].

Overall, I envision that the "fusion" of simulations and machine learning (ML) models (see Fig. 12-2) will unlock a new era in materials modeling—wherein traditional boundaries between physics and empirical models, knowledge and data, forward and inverse predictions, or experimental and simulation data would eventually fade. I hope that the present thesis will modestly contribute to stimulating new developments in that direction.

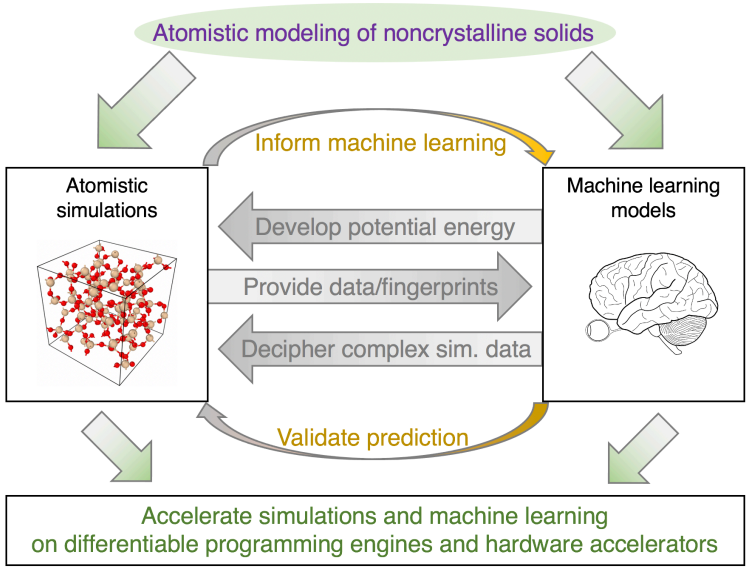


Figure 12-2: Schematic summarizing future opportunities for materials modeling offered by the mutual integration of simulations and machine learning (ML). On the one hand, ML can assist in (i) developing empirical forcefields for accurate and computationally-efficient simulations, (ii) "separating the wheat from the chaff" in large amounts of complex simulation data to gain new insights or generate new knowledge of the underlying physics governing glases, and (iii) accelerating simulations by surrogate machine learning engines. On the other hand, simulation can generate large amounts of high-fidelity data that can be used to train machine learning models, which, in turn, can be validated by simulations. Both simulations and their integration pipeline with ML can be accelerated by leveraging automated differentiable programming engines (e.g., Python JAX) and hardware accelerators (e.g., graphics processing unit (GPU) and tensor processing unit (TPU)). Image adopted from ref. [43]

12.3 References

[1] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Accelerating the Design of Functional Glasses through Modeling, Chem. Mater. 28 (2016) 4267–4277. https://doi.org/10.1021/acs.chemmater.6b01054.

[2] J.C. Mauro, Decoding the glass genome, Current Opinion in Solid State and Materials Science. 22 (2018) 58–64. https://doi.org/10.1016/j.cossms.2017.09.001.

[3] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: A review, Journal of Non-Crystalline Solids: X. 4 (2019) 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[4] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Materials. 4 (2016) 053208. https://doi.org/10.1063/1.4946894.

[5] R. Jose, S. Ramakrishna, Materials 4.0: Materials big data enabled materials discovery, Applied Materials Today. 10 (2018) 127–132. https://doi.org/10.1016/j.apmt.2017.12.015.

[6] K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, N.M.A. Krishnan, M.M. Smedskjaer, C. Hoover, M. Bauchy, Predicting the Young's Modulus of Silicate Glasses using High-Throughput Molecular Dynamics Simulations and Machine Learning, Sci Rep. 9 (2019) 1–11. https://doi.org/10.1038/s41598-019-45344-3.

[7] H. Liu, T. Zhang, N.M.A. Krishnan, M.M. Smedskjaer, J.V. Ryan, S. Gin, M. Bauchy, Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning, Npj Mater Degrad. 3 (2019) 1–12. https://doi.org/10.1038/s41529-019-0094-1.

[8] H. Liu, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Structural percolation controls the precipitation kinetics of colloidal calcium–silicate–hydrate gels, J. Phys. D: Appl. Phys. 52 (2019) 315301. https://doi.org/10.1088/1361-6463/ab217b.

[9] H. Liu, S. Dong, L. Tang, N.M.A. Krishnan, G. Sant, M. Bauchy, Effects of polydispersity and disorder on the mechanical properties of hydrated silicate gels, Journal of the Mechanics and Physics of Solids. 122 (2019) 555–565. https://doi.org/10.1016/j.jmps.2018.10.003.

[10] H. Liu, S. Dong, N.M.A. Krishnan, E. Masoero, G. Sant, M. Bauchy, Long-term creep deformations in colloidal calcium–silicate–hydrate gels by accelerated aging simulations, Journal of Colloid and Interface Science. 542 (2019) 339–346. https://doi.org/10.1016/j.jcis.2019.02.022.

[11] H. Liu, Y. Liu, Z. Zhao, M. Bauchy, S.S. Schoenholz, E.D. Cubuk, End-to-End Differentiability and Tensor Processing Unit Computing to Accelerate Materials' Inverse Design, in: 2020.

[12] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Parameterization of empirical forcefields for glassy silica using machine learning, MRS Communications. (2019) 1–7. https://doi.org/10.1557/mrc.2019.47.

[13] H. Liu, Z. Fu, Y. Li, N.F.A. Sabri, M. Bauchy, Balance between accuracy and simplicity in empirical forcefields for glass modeling: Insights from machine learning, Journal of Non-Crystalline Solids. 515 (2019) 133–142. https://doi.org/10.1016/j.jnoncrysol.2019.04.020.

[14] H. Liu, Y. Li, Z. Fu, K. Li, M. Bauchy, Exploring the landscape of Buckingham potentials for silica by machine learning: Soft vs hard interatomic forcefields, J. Chem. Phys. 152 (2020) 051101. https://doi.org/10.1063/1.5136041.

[15] H. Liu, E. Li, E.D. Cubuk, S.S. Schoenholz, S. Xiao, C. Yang, G. Sant, M.M. Smedskjaer, M. Bauchy, Deciphering a Structural Signature of Glass Dynamics by Machine Learning, Physical Review B. (2021).

[16] H. Liu, S. Xiao, L. Tang, E. Bao, E. Li, C. Yang, Z. Zhao, G. Sant, M.M. Smedskjaer, L. Guo, M. Bauchy, Predicting the early-stage creep dynamics of gels from their static structure by machine learning, Acta Materialia. 210 (2021) 116817. https://doi.org/10.1016/j.actamat.2021.116817.

[17] S. Chmiela, H.E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, Nat Commun. 9 (2018) 3887. https://doi.org/10.1038/s41467-018-06169-2.

[18] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, Nature Materials. 20 (2021) 750–761. https://doi.org/10.1038/s41563-020-0777-6.

[19] C. Massobrio, ed., Molecular dynamics simulations of disordered materials: from network glasses to phase-change memory allyos, Springer, Cham Heidelberg, 2015.

[20] K. Binder, W. Kob, Glassy Materials and Disordered Solids: An Introduction to Their Statistical Mechanics, World Scientific, 2011.

[21] E.D. Cubuk, R.J.S. Ivancic, S.S. Schoenholz, D.J. Strickland, A. Basu, Z.S. Davidson, J. Fontaine, J.L. Hor, Y.-R. Huang, Y. Jiang, N.C. Keim, K.D. Koshigan, J.A. Lefever, T. Liu, X.-G. Ma, D.J. Magagnosc, E. Morrow, C.P. Ortiz, J.M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K.N. Nordstrom, P.E. Arratia, R.W. Carpick, D.J. Durian, Z. Fakhraai, D.J. Jerolmack, D. Lee, J. Li, R. Riggleman, K.T. Turner, A.G. Yodh, D.S. Gianola, A.J. Liu, Structure-property relationships from universal signatures of plasticity in disordered solids, Science. 358 (2017) 1033–1037. https://doi.org/10.1126/science.aai8830.

[22] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[23] E.D. Cubuk, S.S. Schoenholz, J.M. Rieser, B.D. Malone, J. Rottler, D.J. Durian, E. Kaxiras, A.J. Liu, Identifying Structural Flow Defects in Disordered Solids Using

Machine-Learning Methods, Physical Review Letters. 114 (2015). https://doi.org/10.1103/PhysRevLett.114.108001.

[24] S.S. Schoenholz, E.D. Cubuk, D.M. Sussman, E. Kaxiras, A.J. Liu, A structural approach to relaxation in glassy liquids, Nature Physics. 12 (2016) 469–471. https://doi.org/10.1038/nphys3644.

[25] E.O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery, Annual Review of Materials Research. 45 (2015) 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.

[26] T.W. Liao, G. Li, Metaheuristic-based inverse design of materials – A survey, Journal of Materiomics. 6 (2020) 414–430. https://doi.org/10.1016/j.jmat.2020.02.011.

[27] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science. 361 (2018) 360–365. https://doi.org/10.1126/science.aat2663.

[28] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, Journal of Computational Physics. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 12 (2011) 2825–2830.

[30] G. Corliss, C. Faure, A. Griewank, L. Hascoet, U. Naumann, Automatic Differentiation of Algorithms: From Simulation to Optimization, Springer Science & Business Media, 2013.

[31] F. de Avila Belbute-Peres, K. Smith, K. Allen, J. Tenenbaum, J.Z. Kolter, End-to-End Differentiable Physics for Learning and Control, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018: pp. 7178–7189. http://papers.nips.cc/paper/7948-end-to-end-differentiable-physics-for-learning-and-control.pdf.

[32] J. Degrave, M. Hermans, J. Dambre, F. Wyffels, A Differentiable Physics Engine for Deep Learning in Robotics, Front. Neurorobot. 13 (2019). https://doi.org/10.3389/fnbot.2019.00006.

[33] J. Bradbury, R. Frostig, P. Hawkins, M.J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, Q. Zhang, JAX: composable transformations of Python+NumPy programs, 2018. http://github.com/google/jax.

[34] S. Schoenholz, E.D. Cubuk, JAX MD: A Framework for Differentiable Physics, Advances in Neural Information Processing Systems. 33 (2020) 11428–11441.

[35] K. Yang, Y.-F. Chen, G. Roumpos, C. Colby, J. Anderson, High performance Monte Carlo simulation of ising model on TPU clusters, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, Denver Colorado, 2019: pp. 1–15. https://doi.org/10.1145/3295500.3356149.

[36] F. Huot, Y.-F. Chen, R. Clapp, C. Boneti, J. Anderson, High-resolution imaging on TPUs, ArXiv:1912.08063 [Physics]. (2019). http://arxiv.org/abs/1912.08063.

[37] T. Lu, Y.-F. Chen, B. Hechtman, T. Wang, J. Anderson, Large-Scale Discrete Fourier Transform on TPUs, ArXiv:2002.03260 [Cs]. (2020). http://arxiv.org/abs/2002.03260.

[38] Y.E. Wang, G.-Y. Wei, D. Brooks, Benchmarking TPU, GPU, and CPU Platforms for Deep Learning, (2019). https://arxiv.org/abs/1907.10701v4.

[39] H. Liu, Z. Huang, Y. Sun, W. Wang, E.D. Cubuk, S.S. Schoenholz, Z. Zhao, R. Chen, M. Bauchy, Toward algorithm-free molecular dynamics by graph networks, (2021).

[40] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, P.W. Battaglia, Learning to Simulate Complex Physics with Graph Networks, ArXiv:2002.09405 [Physics, Stat]. (2020). http://arxiv.org/abs/2002.09405.

[41] D. Kochkov, J.A. Smith, A. Alieva, Q. Wang, M.P. Brenner, S. Hoyer, Machine learning accelerated computational fluid dynamics, ArXiv:2102.01010 [Physics]. (2021). http://arxiv.org/abs/2102.01010.

[42] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E.D. Cubuk, S.S. Schoenholz, A. Obika, A.W.R. Nelson, T. Back, D. Hassabis, P. Kohli, Unveiling the predictive power of static structure in glassy systems, Nat. Phys. 16 (2020) 448–454. https://doi.org/10.1038/s41567-020-0842-8.

[43] H. Liu, Z. Zhao, Q. Zhou, R. Chen, K. Yang, Z. Wang, L. Tang, M. Bauchy, Present Challenges and Future Developments in Atomistic Modeling of Glasses: A Review, Comptes Rendus Geoscience. (2021).