PhD project

François P. LANDES
Cyril FURTLEHNER

29 mars 2021

Equivariant Graph Neural Networks for amorphous materials

Abstract

Graph Neural Networks (GNNs) are a relatively recent paradigm of Deep Learning, allowing to enforce node-permutation equivariance (re-indexing), a symmetry present in all graphs. Likewise, rotation-equivariant networks (especially GNNs) allow to enforce rotation equivariance at the level of the very architecture (it constrains the choice in aggregation or update steps). The paradigm of (rotation) equivariant GNNs now dominates the field of molecular properties prediction (or ab initio replacements with machine learned energies and forces).

We propose to work on a test case which represents ideal conditions for **studying and designing new equivariant GNNs**: the case of glassy materials. This test case has its own scientific significance for theoretical physics, but also represents a hard problem, so that solving it satisfactorily necessitates the design of extremely expressive, sensitive GNNs. It is somehow different from the more common benchmarks, which usually involve sparse assemblies of atoms (as molecules, or at most an adsorbant), here instead space is densely packed with particles. Predicting glassy dynamics has known a surge of interest recently, and the state of the art is now disputed by different approaches.

Our team has provided a very competitive model based on equivariant GNNs, which outperforms other methods in a range of physical conditions, but not in all of them. The other approaches are very different from ours, so there is room for improvement. Furthermore, we have not yet reach the theoretical upper bound for prediction accuracy. In terms of interpretability, equivariant GNNs offer a better outlook than regular ones. Using the testbed of glassy materials and starting from our existing codebase, the PhD candidate will study equivariant networks, then develop new architectures or training curriculum to improve generalizability and possibly propose strategies to produce interpretable networks, or produce interpretations of the network's output.

Lab: LISN, Université Paris-Saclay
Team: INRIA team TAU (TAckling the Underspecified), within the A & O team,
Advisors: François Landes (francois.landes@upsaclay.fr), Maitre de Conférences, bientôt inscrit en HDR.
Cyril Furtlehner, CR INRIA, HDR.
Head of Lab: Sophie Rosset (rosset@limsi.fr)
Location: France, Gif-sur-yvette (Plateau de Saclay), Batiment 660 (Digiteo building)
École Doctorale: ED STIC, Pôle B.

Financements envisagés :
Co-FUND
Programme blanc GS-ISN
Website: http://lptms.u-psud.fr/francois-landes/

Contents

1 Keywords

Machine Learning, Graph Neural Networks, Equivariant, steerable features, Explainability, Interpretability, Benchmark datasets, data challenge, Machine Learned order parameter, Glassy systems, Supercooled liquids, Dense materials, Statistical Physics.

2 State of the art: SE(3)-equivariant GNN for glassy materials

We provide background information on DNN, GNN and glassy systems in appendix.
For equivariant GNNs, we refer to our own work, [PCL22].

The state of the art has been redefined in late 2022: our own preprint [PCL22], along with a few other independent works [SHSS23, ASF23, JBB23]. It's worth noting that Deepmind's redefinition of the SOTA (from 2020) has been largely surpassed by these 4 independent contributions. **Rather than copy-pasting our own paper here, we refer the reader to [PCL22], which is quite pedagogical as an introduction to equivariant networks, and also presents the problem, our solution, and outlines directions for future work.**

Recent uses of GNNs for disordered materials

A relatively recent review [GY21], **discusses applications of GNNs on solid-state materials**, polymers and by extension, amorphous materials in general. In its conclusions, this review outlines the challenges of the field, quite in line with our own argumentation.

The **timeliness of our proposal is apparent from the number of very recent publications** using some sorts of GNNs on applications related to amorphous or more generally disordered materials. These works, however, are different from our project, since **they regress or classify globally, instead of predicting particle-absed quantities**. For instance, there is a work on phase classification for active materials [DB20]. Another one aims at regressing macroscopic properties of polymers [XFLW+21, Xie] Yet another example is [STL+20], which performs liquid/glass classification (for the whole system, not taking care of the local fluctuations), which is a rather simple task, by design. *Disclaimer: this litterature search has not been updated since late 2021.*

While dense materials are under-studied, the PhD candidate may be interested in attacking more wide-known problems such as molecular properties predictions, such as those defined in the benchmarks OC20, OC22, MD17 or QM9.

3 Goals - Summary

Depending on the interest of the PhD candidate and depending on opportunities, a number of directions can be followed, with various priority order.

3.1 Goal 1: Explainability

Motivation: The machine-learned output of the (successful) GNN, by definition, depends solely on the atomic structure (relative positions of atoms). It is thus a **Machine-Learned *structural order parameter***. Provided the accuracy is very high, this order parameter cannot be discarded as irrelevant by the domain experts. However, it is indeed **hard to interpret and does not constitute, by itself, a progress in basic science**. What would constitute such a progress would be an **explainable** order parameter. On that side, a possible lead would be to train a simpler (interpretable) model to reproduce its parent's output (this is known as *knowledge distillation* [HVD15]). An advantage of our application domain (glasses) is that **one can expect the relationship to be somehow "elegant"** since we are dealing with fundamental physics. In the same time, we know that **it won't be trivial** (otherwise the physics community would have probably cracked the problem, over the last 30-40 years..). This is thus **a good benchmark for building explainable models**.

How to do it: We have several ideas to simplify the network into an explicable one, some of which we have already experimented on, demonstrating no performance decrease. For instance, we have developed a variation where channels are associated to bond types, showing that this does not decrease performance, while making the network's operations more interpretable. Another idea is to get rid of the MLPs in the message-building step, while increasing channel number (this consequently increases the linear layer width at channel mixing time, but this operation is somehow more interpretable than MLPs, furthermore one could consider sparsifying this linear layer). *Disclaimer: for lack of time, I don't develop further here, but this can be discussed live with the interested applicants*

A number of other directions for interpretability can be discussed. This topic is actively researched at TAU, so there are resources around, to discuss and design new methods (Michèle Sebag in particular is active on this topic).

3.2 Goal 2: Better understanding of equivariant neural networks

Starting from our testbed case of glassy materials, we may attempt new experiments in the spirit of [JBM+23] to study the expressivity of equivariant architectures. It is unclear how exactly some recent architectures are equivariant, and what is the role of type order vs network depth and channel number.

This of course is in interplay with the previous goal.

3.3 Goal 3: performance

We have already designed an equivariant GNNs that tackles the problem of amorphous materials However, results can still be improved.

For that, we will need to tap into the existing equivariant GNN toolbox, but beyond that, most likely, to **extend this (continuously growing) toolbox**.

3.4 Goal 4: Dissemination, Benchmarks and Data challenge

Dissemination

The successful GNN would be **properly packaged, documented, possibly pre-trained** on a number of standard glassy mixtures, to **foster the use and understanding of ML in the fundamental physics community**. The open code would be an invitation for collaborative work on related application domains (various mixtures possibly needing small adaptations to handle system-specific features).

Benchmarks

In general, the **physics community lacks a culture of sharing benchmark dataset**, reproducible baselines. This project could be a push in the right direction, showing the benefits of a more open culture (for data, and for code). Thus, a side-benefit of our project would be the **release of a benchmark of glassy materials datasets**, aiming at node regression, for GNNs (or other architectures) to be compared on.

Data challenge

We may also note that the Auto-ML track of the 2020 KDD cup edition https://www.kdd.org/kdd2020/kdd-cup was focused on "AutoML for Graph Representation Learning", that is, Auto-ML for Graph-based datasets https://www.automl.ai/competitions/3, meaning that participants had to provide algorithms that could easily adapt to variable graph-based data. A **similar challenge may be organized around amorphous solids data**, to foster creativity in the design of amorphous-materials specialized GNNs. More precisely, an **Auto-ML challenge on amorphous material data** could help finding GNNs that can learn representations for all materials (ideally, performing **transfer learning between different materials**).

More humbly, one of us has already presented a problem similar to this project in the context of the **Data Challenge organized by Pr. S. Mallat, at the Collège de France**: Challenge Data 2018.

https://challengedata.ens.fr/en/challenge/46/solve_a_major_physics_problem_by_finding_how_to_predict_the_dynamics_of_glass-forming_liquids.html

At the time, the definition of labels was rather simplistic, which made the task quite unpractical to address. Also, a single material was provided, as we were far from speculating about transfer learning between different materials.

3.5 Goal 5: Dimensional reduction

An avenue for future research is to build coarse-grained models of glassy physics, in the spirit of [**?**] (some of whom are former collaborators of us). This is still an ML task in that it can be seen as dimensional reduction of large time series of 3D point cloud data. This perspective on the problem being, by the way, absolutely new to physicists of that field (and, to my knowledge, new in general, for solid materials at least - for liquids, Navier-Stokes is a very old dimensional reduction "model").

Disclaimer: for lack of time, I don't develop further here, but this can be discussed live with the interested applicants

Another application could be that of quasi static amorphous shear, in which isotropy is broken, since shear imposes a preferred direction. This could be attacked with equivariant Neural Networks. We know Sylvain Patinet (Jussieu) who works on this topic (on the physics side). A collaboration on this topic could start.

Motivation: The study of glasses and in particular of their dynamics is a **major subject in statistical physics, which has attracted the attention of many physicists for more than 30 years now**. It is a major problem because there are signs indicating the existence of a **new form of phase transition** (a central and fruitful concept in statistical physics), and links can be made with other disciplines. For example the increasingly slow exploration of the disordered energy landscape, a central point of the phenomenology of glasses, corresponds to the search for a near-optimal solution of an NP-hard problem by Monte Carlo methods (approximately speaking, Newton's equations of motion). A central question is to understand how "dynamic heterogeneities" contribute to the displacements in the glasses. Indeed, it has been observed that in a large glassy system, activity is concentrated on restricted areas during restricted periods of time, i.e. the dynamics is heterogeneous, spatially and temporally.

A **key question** that one of us has been studying for some time is to understand the avalanche dynamics characteristic of dynamic heterogeneities in supercooled liquids, i.e. **how activity nucleates and propagates**. In particular, we wish to understand the respective importance of facilitation (how local activity facilitates the activation of dynamics in the neighborhood) vis-à-vis that of **local structure** (*i.e.* the input of the GNN), in the birth, propagation and death of avalanches. This question is crucial because it is the progressive disappearance of dynamic heterogeneities with cooling that makes glasses

be solid materials, or conversely it is the appearance of these heterogeneities with heating that allows the (difficult) exploration of the disordered energy landscape, characteristic of NP-hard problems.

In the practical sense, understanding glasses would allow to **better control and possibly to better design materials of this extremely broad class (glasses)**. In terms of hazard control, solving this fundamental question would allow **better control over unknown phenomena**: for instance, when we mix **radioactive waste into a glassy material, how will this glass age** in 100.000, or 1.000.000 years ?

3.6 Other possible developments

An additional application for dense-materials able GNNs would be to infer properties for proteins *inside a solvent*. A large part of the protein-studying community usually ignores the existence of the bath in which a protein (or polymer) necessarily lives in. At best, the solvent is present in the Molecular Dynamics simulations that are used as training set, but the solvent is then considered as an average bath, playing no crucial role. However, this view is disputed [BFHH+16] and the role of the solvent can be crucial, so that **considering the full protein+solvent system may be absolutely necessary**. The **models we aim to develop would naturally be able to tackle this kind of data**, as they naturally deal with large, dense assemblies of heterogeneous atoms, in which tiny changes in the atomic positions can play a crucial role (as is known to be the case in the protein-ligand interaction, for instance).

In glasses, other properties may be used as target labels, such as the existence and **location of Two-Level systems (TLS) in ultra-low temperature glasses** [KRZ20].

4 Supervision, environment

4.1 PhD roadmap

The successful candidate's goal will be to understand and **critically assess the performance of past GNNs, and design new ones**, taking into account the specifics of the problem at hand. Depending on the interests and abilities of the PhD student, this task may be attacked more or less directly.

At first, the PhD student will be expected to get some knowledge about glassy dynamics, from the physics literature. In the same time, they should **dive into the equivariant GNNs literature**, as outlined in this project. The literature review should be extensive enough so as to allow **a critical view of the recent literature to emerge**.

Second, the PhD student should be able to reproduce past experiments: setting up a sound task (input and outputs well-defined and well-motivated), they should compare the relevant state-of-the-art architectures. This would also be an opportunity to have a more concrete sense of each design choice. From this review of existing algorithms, **a sense of the critical design choices should emerge**.

Beating the state-of-the-art methods, or challenging them with innovative approaches (*e.g.* successfully using expert knowledge) will open the way for **publications in International Conferences or Journals, primarily in the Machine Learning** community, but also in the application domain (glassy physics).

Simultaneously, and depending on the results, the PhD student could at least **produce an open dataset of amorphous materials**, to benchmark the algorithms on a sound basis. Possibly, we could organize an **Auto-ML challenge focused on amorphous materials.**

Lastly, on the most performing architectures (be they from us or not), the student will work on **building explainable or transparent models**, either directly on the task, or from the learned representations. Related to this line of research would be to test the ability of the network to perform **transfer learning**, which would answer the question "did the network learn some generic representation, which remains relevant across multiple amorphous materials?"

4.2 Expected abilities of the PhD candidate

The PhD application is open to candidates from a pure ML (Machine Learning) background, but also to Statistical Physicists with a strong ML background.
In any case, the candidate is expected to be **fully proficient in (numerical) python, and/or C++**

(at least one imperative, structured language), and **strong proficiency is expected in mathematics** (algebra, calculus, probability, etc).
A **strong proficency in Machine Learning**, and a strong interest in it, is needed. The ideal candidate will have **some prior knowledge of Deep Learning**, and of a standard library (e.g. PyTorch). Some knowledge of basic **Statistical Physics** (statistical equilibrium, entropy, etc) would be a great plus, since it's helpful to understand the underlying application problem. The ideal candidate would have some more advanced knowledge of Statistical Physics (phase transitions, order parameter, disordered media, out-of-equilibrium dynamics, etc).

4.3 Supervising team

François Landes

François Landes is *Maître de Conférences* in the Inria Saclay project-team TAU since 2018. With an applied mathematics Engineer degree and an M2 in Theoretical Physics, he graduated from Paris Saclay in Statistical Physics in 2014. After 2 years of Postdoc at ICTP (Trieste, Italy), where he diversified on various subjects (among which Machine Learning), he went on to join the Simons Collaboration for another 2 years of a postdoc between Paris and Philadelphia. He has already worked and is still working on the topic of glassy dynamics [LBD+19, ALBB20, CLB+21] and is thus knowledgeable about glassyness or amorphous materials in general. Furthermore, he monitored a short internship on the topic in 2019, which helped to clear out uncertainties about which labels to train on, revealing how this is crucial to the task (to put it shortly: a random result is harder to predict than a more deterministic one). Currently, François monitors a PhD student on the topic of GNNs for glassy materials (Francesco Saverio Pezzicoli).

François Landes would be the main *co-supervisor* of the PhD student, the lead adviser.

Cyril Furhlehner

Cyril Furtlehner: HDR, *chargé de recherche* in the Inria Saclay project-team TAU. Graduated in theoretical physics from Paris VI University in 1997. One year as post-doc in the Max Plank institute of Heidelberg and two years at the university of Oslo working on quantum disordered systems. Joined the Inria team PREVAL in 2001 to work on stochastic processes and more applied research topics like the self organization of a fleet of automated vehicles. Permanently hired in 2007 in the Inria TAO team, with scientific activities at the interface between statistical physics, machine learning and optimization.

4.4 Broader scientific Environment

4.4.1 Johannes Brandstetter

Johannes Brandstetter came to visit TAU in 2023, we have a friendly relationship since then. He has been working on equivariant networks with the pioneers (e.g. Max Welling) in Amsterdam. It would be possible to organize visits to his current lab (he recently changed position).

4.4.2 Simons Collaboration

One of us (F. Landes) is an alumni of the Simons Collaboration on *Cracking the Glass Problem* (apologies for the pun), an international collaboration of researchers working on the glass problem, with funding from the Simons Fundation (10 M$ over 4 years, renewed for an additional 4M$ over 3 more years). This provides us with a number of past collaborators, so that the expert knowledge we may need will be available.

This also provides us access to an expert community which is rather critical of the potential benefits of ML, and still rather uneasy with the methodological fundamentals of the field. This situation creates an opportunity to challenge the existing views and foster the development and use of ML in new contexts.

4.5 Local Environment: the team and the lab

The TAU team (TAckling the Underspecified, INRIA team) and the A&O team, to which all three supervisors belong to, is specialized in Machine Learning, with a number of members being keen on

Deep Learning fundamentals and their applications. A growing number of members became interested in GNNs over the last year, so that the PhD student would fit in quite naturally in the team.

The infrastructure (clusters, knowledge base) is readily available to the team members, and is adapted to welcome the PhD student.

References

[ALBB20] Francesco Arceri, François P. Landes, Ludovic Berthier, and Giulio Biroli. Glasses and aging: A Statistical Mechanics Perspective. *arXiv:2006.09725 [cond-mat]*, October 2020. arXiv: 2006.09725.

[ASF23] Rinske M. Alkemade, Frank Smallenburg, and Laura Filion. Improving the prediction of glassy dynamics by pinpointing the local cage. *The Journal of Chemical Physics*, 158(13):134512, 04 2023.

[BBX+16] Ramin Bostanabad, Anh Tuan Bui, Wei Xie, Daniel W. Apley, and Wei Chen. Stochastic microstructure characterization and reconstruction via supervised learning. *Acta Materialia*, 103:89–102, January 2016.

[BFHH+16] Marie-Claire Bellissent-Funel, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David van der Spoel, Yao Xu, and Angel E Garcia. Water Determines the Structure and Dynamics of Proteins. *Chemical Reviews*, 116(13):7673–7697, July 2016.

[BHB+18] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*, October 2018. arXiv: 1806.01261.

[BKGB+20] 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, and P. Kohli. Unveiling the predictive power of static structure in glassy systems. *Nature Physics*, 16(4):448–454, April 2020.

[BMAM+20] Emanuele Boattini, Susana Marín-Aguilar, Saheli Mitra, Giuseppe Foffi, Frank Smallenburg, and Laura Filion. Autonomously revealing hidden local structures in supercooled liquids. *Nature Communications*, 11(1):5479, December 2020.

[BPL+16] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction Networks for Learning about Objects, Relations and Physics. *arXiv:1612.00222 [cs]*, December 2016. arXiv: 1612.00222.

[CDY+18] Ahmet Cecen, Hanjun Dai, Yuksel C. Yabansu, Surya R. Kalidindi, and Le Song. Material structure-property linkages using three-dimensional convolutional neural networks. *Acta Materialia*, 146:76–84, March 2018.

[CG] Sathya Chitturi and Xiaotong Gui. Deep Surrogate Models for Atomic Systems. page 6.

[CLB+21] Rahul N. Chacko, François P. Landes, Giulio Biroli, Olivier Dauchot, Andrea J. Liu, and David R. Reichman. Elastoplasticity Mediates Dynamical Heterogeneity Below the Mode-Coupling Temperature. *arXiv:2103.01852 [cond-mat, physics:physics]*, March 2021. arXiv: 2103.01852.

[CYZ+19] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, May 2019. arXiv: 1812.05055.

[DB20] Austin R. Dulaney and John F. Brady. Machine Learning for Phase Behavior in Active Matter Systems. *arXiv:2011.09458 [cond-mat, stat]*, November 2020. arXiv: 2011.09458.

[GJ19] Hongyang Gao and Shuiwang Ji. Graph U-Nets. *arXiv:1905.05178 [cs, stat]*, May 2019. arXiv: 1905.05178.

[GSR+17] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212 [cs]*, June 2017. arXiv: 1704.01212.

[GY21] Weiyi Gong and Qimin Yan. Graph-based deep learning frameworks for molecules and solid-state materials. *Computational Materials Science*, page 110332, March 2021.

[HSD+21] Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec, Devi Parikh, and C. Lawrence Zitnick. ForceNet: A Graph Neural Network for Large-Scale Quantum Calculations. *arXiv:2103.01436 [cs]*, March 2021. arXiv: 2103.01436.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, March 2015. arXiv: 1503.02531.

[JBB23] Gerhard Jung, Giulio Biroli, and Ludovic Berthier. Predicting dynamic heterogeneity in glass-forming liquids by physics-inspired machine learning. *Phys. Rev. Lett.*, 130:238202, Jun 2023.

[JBK+20] Devendra K. Jangid, Neal R. Brodnik, Amil Khan, McLean P. Echlin, Tresa M. Pollock, Sam Daly, and B. S. Manjunath. 3DMaterialGAN: Learning 3D Shape Representation from Latent Space for Materials Science Applications. *arXiv:2007.13887 [cs, eess]*, July 2020. arXiv: 2007.13887.

[JBM+23] Chaitanya K Joshi, Cristian Bodnar, Simon V Mathis, Taco Cohen, and Pietro Lio. On the expressive power of geometric graph neural networks. 2023.

[JJS] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N Schmidt. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. page 11.

[KKK+20] QHwan Kim, Joon-Hyuk Ko, Sunghoon Kim, Nojun Park, and Wonho Jhe. Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. *arXiv:2012.08194 [cs]*, December 2020. arXiv: 2012.08194.

[KMB+16] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, August 2016.

[KRZ20] Dmytro Khomenko, David R. Reichman, and Francesco Zamponi. Relationship between two-level systems and quasi-localized normal modes in glasses. *arXiv:2012.07998 [cond-mat]*, December 2020. arXiv: 2012.07998.

[LBD+19] François P. Landes, Giulio Biroli, Olivier Dauchot, Andrea J. Liu, and David R. Reichman. Attractive vs. truncated repulsive supercooled liquids : dynamics is encoded in the pair correlation function. *arXiv:1906.01103 [cond-mat]*, June 2019. arXiv: 1906.01103.

[LYB+18] Xiaolin Li, Zijiang Yang, L. Catherine Brinson, Alok Choudhary, Ankit Agrawal, and Wei Chen. A Deep Adversarial Learning Methodology for Designing Microstructural Material Systems. In *Volume 2B: 44th Design Automation Conference*, page V02BT03A008, Quebec City, Quebec, Canada, August 2018. American Society of Mechanical Engineers.

[NBC17] Andrea Ninarello, Ludovic Berthier, and Daniele Coslovich. Models and Algorithms for the Next Generation of Glass Transition Studies. *Physical Review X*, 7(2), June 2017.

[PCL22] Francesco Saverio Pezzicoli, Guillaume Charpiat, and François P Landes. Se (3)-equivariant graph neural networks for learning glassy liquids representations. *arXiv preprint arXiv:2211.03226*, 2022.

[SHSS23] Hayato Shiba, Masatoshi Hanai, Toyotaro Suzumura, and Takashi Shimokawabe. BOTAN: BOnd TArgeting Network for prediction of slow glassy dynamics by machine learning relative motion. *The Journal of Chemical Physics*, 158(8):084503, 02 2023.

[SSK+18] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, June 2018.

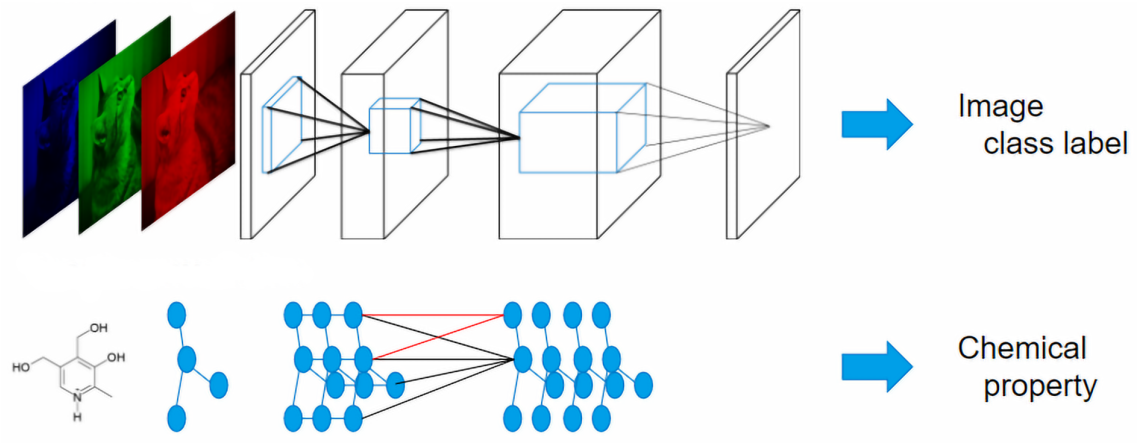

Fig. 1: CNNs (up) compared with GNNs (bottom). In the GNN case, only one convolution operation has been sketched, focusing on the node update of the central node (black), and of the role of edge features for another node (red lines).

[STL+20] Kirk Swanson, Shubhendu Trivedi, Joshua Lequieu, Kyle Swanson, and Risi Kondor. Deep learning for automated classification and characterization of amorphous materials. *Soft Matter*, 16(2):435–446, 2020.

[VCC+18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *arXiv:1710.10903 [cs, stat]*, February 2018. arXiv: 1710.10903.

[XFLW+21] Tian Xie, Arthur France-Lanord, Yanming Wang, Jeffrey Lopez, Michael Austin Stolberg, Megan Hill, Graham Michael Leverick, Rafael Gomez-Bombarelli, Jeremiah A. Johnson, Yang Shao-Horn, and Jeffrey C. Grossman. Accelerating the screening of amorphous polymer electrolytes by learning to reduce random and systematic errors in molecular dynamics simulations. *arXiv:2101.05339 [cond-mat]*, January 2021. arXiv: 2101.05339.

[Xie] Tian Xie. Deep Learning Methods for the Design and Understanding of Solid Materials. page 145.

A Background: Deep Learning and Graph Neural Networks (GNNs)

Here we provide basic information about GNNs and how they relate to CNNs. This is a very simple summary.

In the last decade (since 2012), Deep Neural Networks (DNN) have become an iconic success of Machine Learning, renewing the interest for the subject, to say the least. Particular attention has been paid to Convolutional Neural Networks (CNN), which are very well suited to visual data, and have yielded impressive results. However, this recent success also comes from the **automation** of the algorithms (automated differentiation performed by libraries), the availability of **large datasets** and the development of **GPGPUs**. These last three features can be put at use in the more recent **Graph Neural Networks (GNNs)**, [GSR+17, BHB+18]. The comparison of Convolutional Neural Networks (CNNs) with GNNs is sketched in Fig. 1.

GNNs are able to handle graph-based data, and in particular, the family of so-called **Graph Convolutional Networks (GCN) can perform convolutions** adapted to a variable local node structure. The underlying key idea is to **perform local operations** (as multiplying a neighborhood by a local filter), that can **adapt to the variety of nodes' geometries** (variable degree, as opposed to the constant geometry seen in an image). The weights are typically updated by **message passing routines** [GSR+17], and shared across the graph (by definition, in a convolution, the same operation is applied everywhere, with the exact same weights). This allows GNNs to **generalize**, from the intrinsic relationships between nodes, **beyond the set of observed graphs**, efficiently performing various tasks such as

classification (of graphs, or of their nodes) or multi-variate regressions (again, of global graph quantities, or of local, node-level ones).

For concreteness, we now briefly **sketch the Message-Passing Neural Network (MPNN) paradigm, which is central to GNNs**. We borrow from the seminal [GSR+17], which unified various GNNs approaches under this MPNN paradigm, in a concise manner. G is an undirected graph with node features x_v and edge features e_{vw}. The **forward pass** has two phases, a message passing phase and a readout phase. The **message passing phase runs for T time steps (the more steps, the deeper the network)** and is defined in terms of "message functions" M_t and vertex update functions U_t. During the message passing phase, hidden states h_v^t at each node in the graph are updated based on messages m_v^{t+1} according to

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{1}$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{2}$$

where in the sum, $N(v)$ denotes the neighbors of v in graph G. The readout phase computes a feature vector (for instance for the whole graph) using some readout function R according to

$$\hat{y} = R(\{h_v^T \mid v \in G\}). \tag{3}$$

The message functions M_t, vertex update functions U_t, and readout function R **are all learned differentiable functions**. R operates on the set of node states and must be invariant to permutations of the node states in order for the MPNN to be invariant to graph isomorphism. **Many GNNs can be cast into this paradigm, their difference lying only in the form of the functions** M_t, U_t, R [GSR+17]. An example choice for this learned functions is to take:

- $M(h_v, h_w, e_{vw}) = (h_w || e_{vw})$ where $(.||.)$ denotes concatenation;
- $U_t(h_v^t, m_v^{t+1}) = \sigma(H_t^{\deg(v)} m_v^{t+1})$, where σ is the sigmoid function, $\deg(v)$ is the degree of vertex v and H_t^N is a **learned** matrix for each time step t and vertex degree N;
- $R = f\big(\sum_{v,t} \text{softmax}(W_t h_v^t)\big)$, where f is a neural network and W_t are **learned** readout matrices, one for each time step t.

The versatility of the graph structure has drawn a lot of **attention to GNNs in the past 2-3 years**. Applications range from **relational learning** [BPL+16], to **social networks** analysis, where one can classify parts of the graph [GJ19] to **biochemistry**, for predicting **proteins** interactions [VCC+18, KKK+20] (see [KMB+16] for a seminal paper presenting the change of paradigm), to physics, for **quantum properties** prediction (on the famous QM9 dataset and its derivatives) [GSR+17, SSK+18, CYZ+19, JBK+20, JJS] or for approximate **effective forces** computations [CG, HSD+21] (much faster than explicit physics-based Density Functional Theory computations can). In physics, more specifically, GNNs have been applied to the analysis and understanding of **dense assemblies of atoms** (solids), for predicting either their electronic structure (band gaps, etc) or local atomic properties properties (atomic energies), or more generally, some quantum properties. Results are quite impressive, as trained models are able to generalize to unseen materials, provided they share the same atomic components as those seen in the train set. This kind of transfer learning manifests the ability of the networks to capture the underlying physics, which is generic across materials (provided the basic constituents are unchanged, of course).

However, these examples providing predictions that **all relate to crystalline structures** (see [JBK+20] for a review of these applications), *i.e.* for cases where the atoms (nodes of the graphs) are periodically arranged in space, and for which the superimposed graph actually matches a physical entity, *i.e.* the bonds between atoms [1] . This field has its own difficulties (the crystal being infinite, the electronic bonds being delocalized over the whole material, etc), however, there is some intrinsic comfort and simplicity in the regularity of the spatial organization of atoms. This implies that crystalline-able architectures will a priori not adapt to amorphous geometries.

[1] The field of microstructures prediction deals with non-crystalline materials, but is substantially different, as it deals with 2D images obtained by experiments, aiming at reconstructing the full image or predicting global material properties [BBX+16, CDY+18]. The approaches that we know of have been limited to rather simple ML techniques, with the exception maybe of [LYB+18], using GANs (but no GNN).

B An ideal benchmark: glassy systems

Here we motivate the application domain (glassy systems) and provide background on the problem to be solved.

B.1 Application domain: Amorphous materials and glasses

We detail this part rather extensively to let the CS-trained reader **appreciate the suitability of the chosen application domain**.

In fundamental physics, a crucial and unsolved problem is that of understanding the behavior of **structural glasses** (also called **glassy liquids**, or **supercooled liquids**). For these materials, there is **no known function that can infer the local dynamical state** (liquid/solid or locally mobile/immobile) from the local geometrical arrangement of the particles (their local *structure*). This task is an obvious goal for Machine Learning, and in particular, for GNNs. We now **present the topics of amorphous materials** in general and glassy systems in particular in more detail.

Glasses or glassy materials are those for which the viscosity or any other characteristic time increases very quickly (about 13 orders of magnitude) when the temperature varies by only a few tens of percent, **without any *obvious* change in the geometrical structure** of their elementary constituents (although a subtle change does occur!). This increase in viscosity means that **from a practical standpoint, the supercooled liquid is a solid, and is then called a glass**. This behavior is to be compared with that of crystalline solids, for which the liquid/solid transition is accompanied by a very clear change of the structure, which passes from an amorphous state to a very structured state, with long range order.

Many materials are amorphous: window glass (silica oxide), but also basalt (and many other rocks), most plastics (assemblies of polymers, for which it is very difficult to crystallize), but also pastes, gels, creams ("soft matter"). In these materials, the various physical measures that can be considered (rigidity, heat capacity) can have a long memory of the previous macroscopic conditions, provided that one cooled quickly enough to "fall out of equilibrium". This memory is particularly evident insofar as physical properties such as ductility/fragility or mechanical strength **depend on their preparation mode**: for instance, this is why tempered glass (*verre trempé*) (or quenched steel, *acier trempé*) are different from their regular versions (cooled progressively) [2]. The existence of this effective "memory" proves that **the internal organization of glasses is non-trivial**: although apparently amorphous and equivalent, two glasses of the same composition, at the same temperature, pressure, etc., but prepared differently, have different macroscopic mechanical properties, **due to their different microscopic structural organizations** (but the quantitative measure of this difference has eluded physicists fro 30 years, and still does today). This **very subtle dependence of physical properties on the structural organization of the atoms makes this prediction task particularly challenging, and explains why crystalline-optimized methods fail on amorphous materials**.

B.2 Why glassy systems are an ideal benchmark

Here **we provide evidence** that glassy systems are an **ideal benchmark for designing new Graph Neural Networks** with strong generalization power.

The problem we have at hand is a rather well-posed problem. **We can generate Molecular Dynamics simulation data at will**, which we call **trajectories**, as they consist in the time series of the positions (and velocities) of the particles inside the simulation box, $\{\mathbf{r}_i(t), \mathbf{v}_i(t)\}_{t\in[0,T],i\in[|1,N|]}$. Some constants are known, like the particles types (at least two), the simulation box size, and the ambient Temperature. From the particle types and relative position, we know exactly the interaction energy of each pair of atoms (or equivalently, the force they mutually apply on each other). The target label may be **some measure of mobility at current or future times**, that we may call $m_i(t+\tau)$, for each particle i, at time t (setting once and for all the time gap τ).

To be concrete, we recall that **the task at hand is to predict**, for a given snapshot of the system, *i.e.* for a fixed time t, the mobility $y = \{m_i(t+\tau)\}_i$ of all atoms, from the input $X = \{\mathbf{r}_i(t)\}_{i\in[|1,N|]}$. So

[2] Steel is a polycrystalline material, but the logic is similar

we perform node-features regression, or multi-variate regression, if we stack mobility at various times, $\mathbf{y} = y_\tau$. The number of particles approximately lies in $N \in [1000, 64000]$.

A strong point in favor of structural glasses is that we know from first principles that the label can be made to be **either fully deterministic** [3] or, for an appropriate choice of mobility measure [LBD+19], the level of structure-dynamics correlation can be computed, *i.e.* **we can compute the upper bound of the accuracy** [4]. Simply said, the data is **noiseless** (no measuring error) and **contains the signal** to rely on. A nice take-away is that the **difficulty or predictability can be tuned**: at high temperatures, most of the signal is in the velocities, making a purely structural description $X = \{\mathbf{r}_i(t)\}_{i\in[|1,N|]}$ irrelevant for prediction. At lower temperatures, the structure becomes more prevalent, and the relationship between $X = \{\mathbf{r}_i(t)\}_{i\in[|1,N|]}$ and y is more deterministic. There is a lower limit to the temperatures we are able to equilibrate to, although it has been lowered by recent advances in simulation techniques [NBC17].

Simultaneously, we know from the start that **the function we want to model,** $y = f_\theta(X)$ **is quite intricate**, since it is known to depend on **the full relative positions of patches of hundreds of particles** (at least about 500 of them at the lowest temperatures), and is –probably– rather sensitive to the details of these relative positions (how exactly, this is the open problem). This is obviously a non-trivial problem, as can be seen by considering the sheer number of configurations accessible to the patch of $n \sim 500$ particles. This patch of particles around a target particle is embedded in a $3n$ dimensional space, but is constrained to actually live on a much smaller-dimensional manifold (smaller intrinsic dimension). Albeit finite, this configuration space is simply too large to be exhaustively sampled.

If the problem was to become too easy (we are far from that point today), we may simply consider assemblies of polymers (or dimers to start with) instead of assemblies of single atoms, and see whether strong generalization remains.

[3] Labels are deterministic if in the definition of the task, we include the velocities in the input $X = \{\mathbf{r}_i(t), \mathbf{v}_i(t)\}_{i\in[|1,N|]}$

[4] For this latter upper-bound to be computed, one needs to perform expensive numerical simulations, using the so-called iso-configurational ensemble (making clones of the original configuration and running it using various velocities samples).