



---

# The Search for Novel Mesoscale Materials

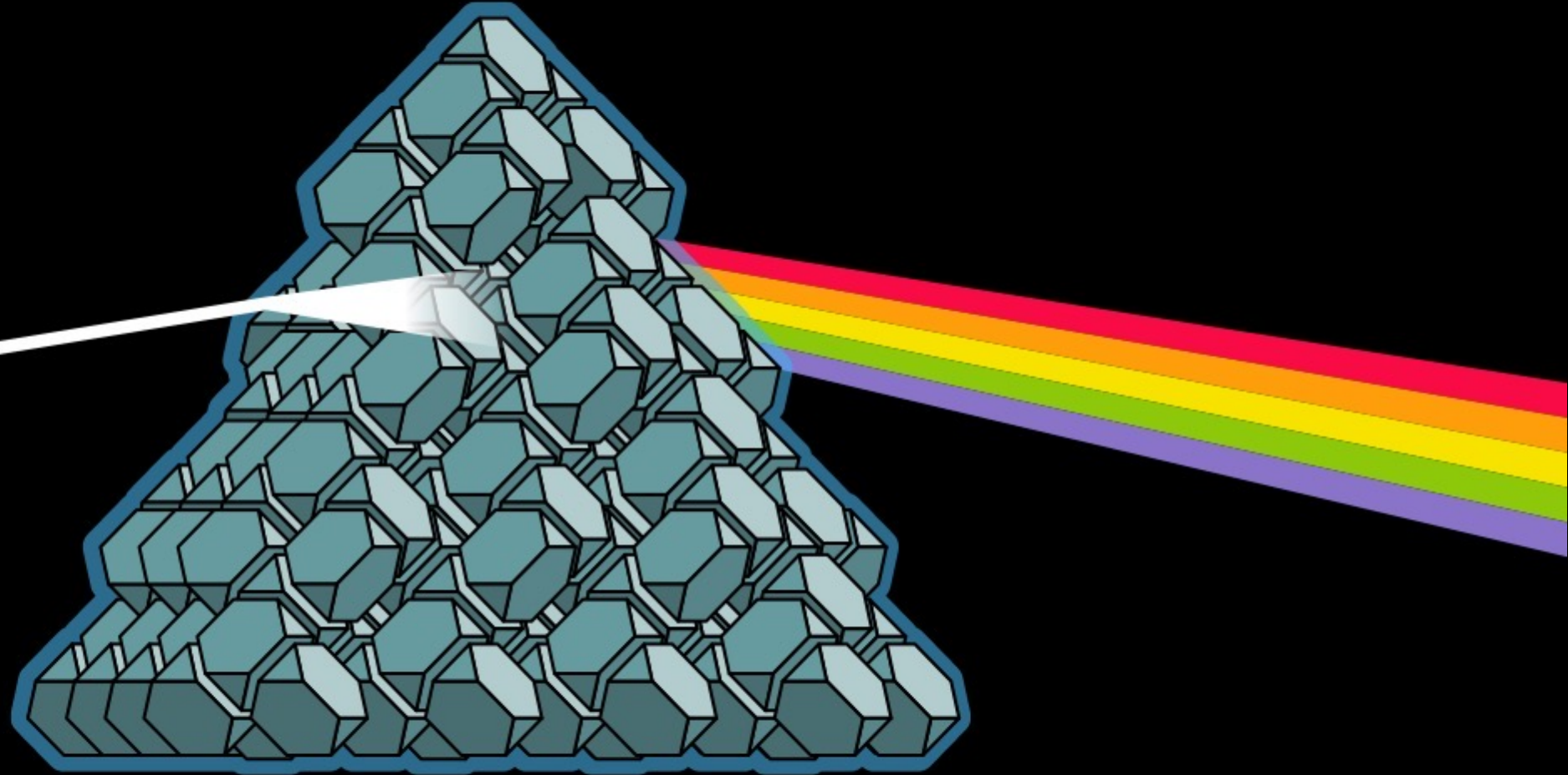
---

**Rose K. Cersonsky**

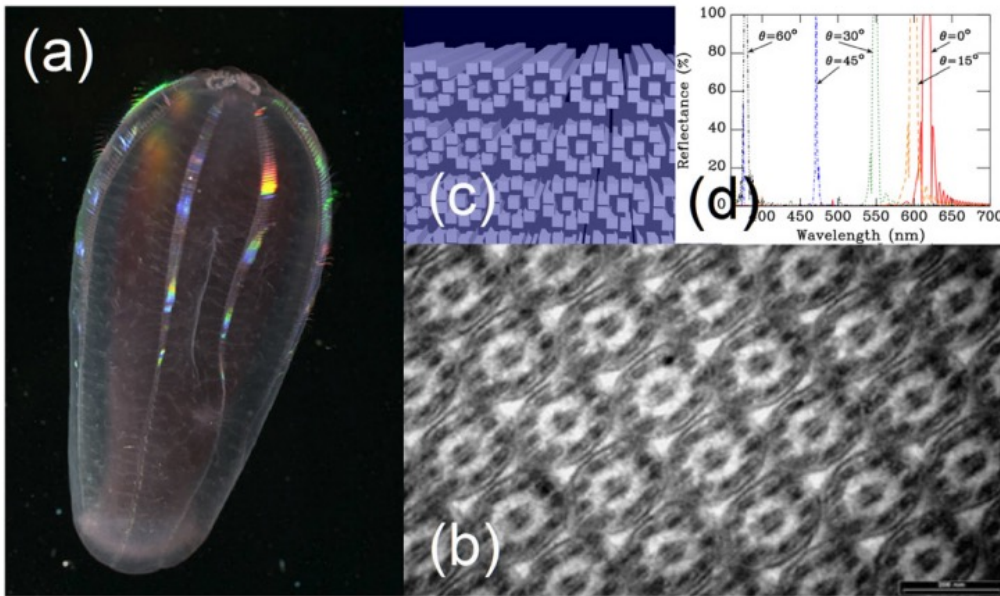
Laboratory of Computational Science and Modeling  
(COSMO)

École Polytechnique Fédérale de Lausanne (EPFL)

Lausanne, Switzerland

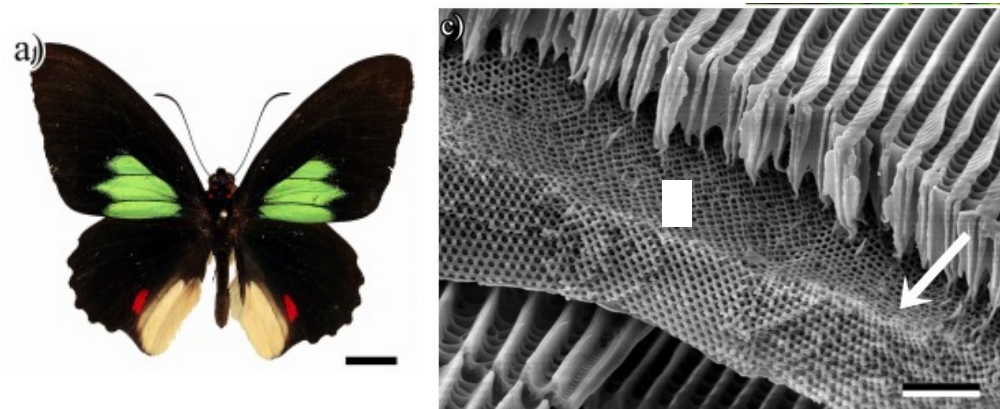


## Designing Nanoparticles for the Self-Assembly of Novel (Photonic) Materials



**Optical properties of the iridescent organ of the comb-jellyfish *Beroë cucumis* (Ctenophora)**

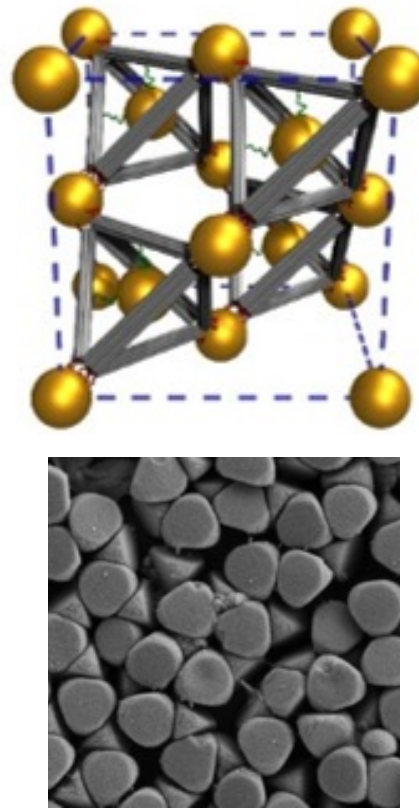
Victoria Welch, *et al.*  
 Phys. Rev. E 73, 041916 2006



**Optical properties of gyroid structured materials: from photonic crystals to metamaterials**

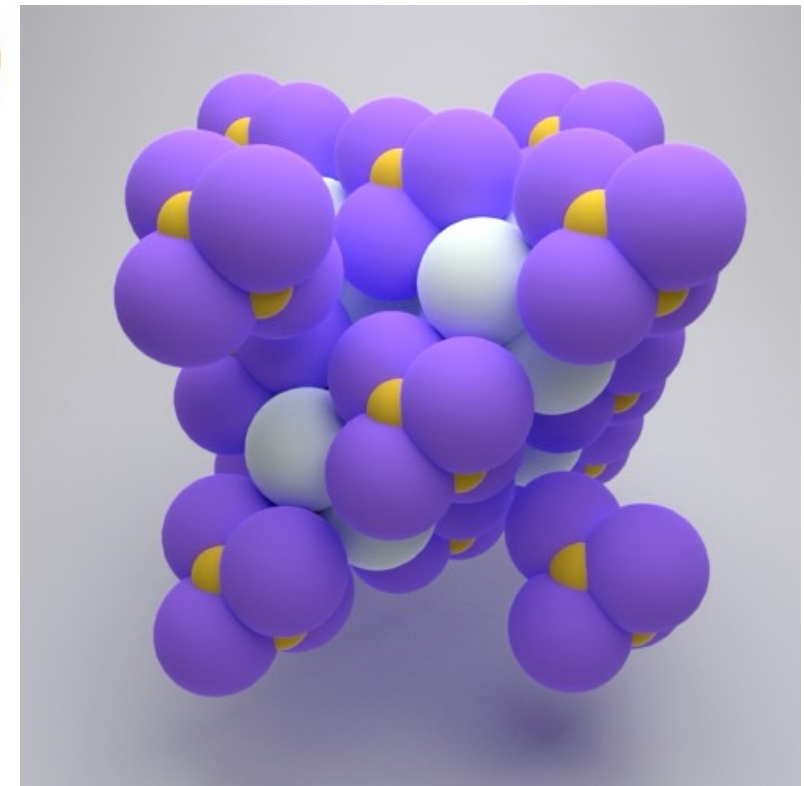
James A. Dolan, *et al.*  
 Advanced Optical Materials 3 (1), 12-32

November 7, 2021



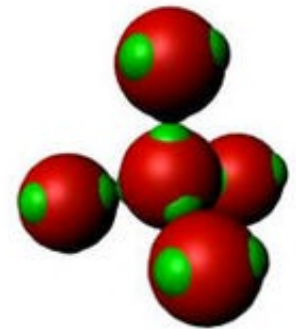
Colloidal crystals with diamond symmetry at optical lengthscales  
 Yifan Wang, *et al.*  
 Nature Comm. 8, 14173 (2017)

Entropy driven assembly of truncated colloidal tetrahedra into diamond structure  
 Zhe Gong, *et al.*



Colloidal Diamond  
 He, M., *et al.*  
 Nature 585, 524-529 (2020).

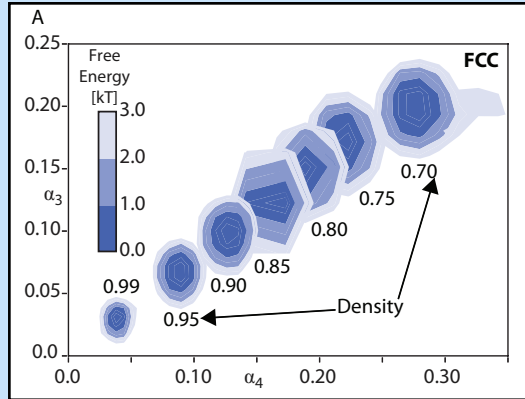
Diamond family of nanoparticle superlattices  
 W. Liu, *et al.*  
 Science 351, 582-586 (2016).





## Relevance of packing to colloidal self-assembly.

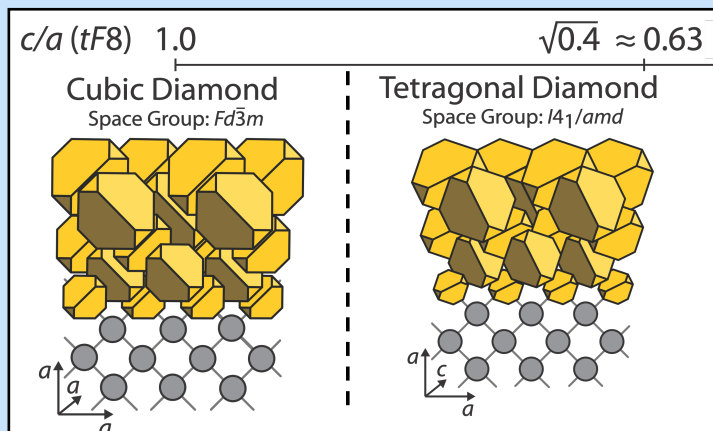
Cersonsky, R. K., van Anders, G., Dodd, P. M., & Glotzer, S. C. (2018). *Proceedings of the National Academy of Sciences*, 115(7), 1439-1444.



- Pauling's packing rules are **not** a causal mechanism for nanoparticle self-assembly
- Using the Digital Alchemy framework, I showed that **adding small imperfections** to nanoparticle shapes would better stabilize nanocrystals

## Pressure-Tunable Photonic Band Gaps in an Entropic Colloidal Crystal

Cersonsky, R. K., Dshemuchadse, J., Antonaglia, J., van Anders, G., & Glotzer, S. C. (2018). *Physical Review Materials*, 2(12), 125201.



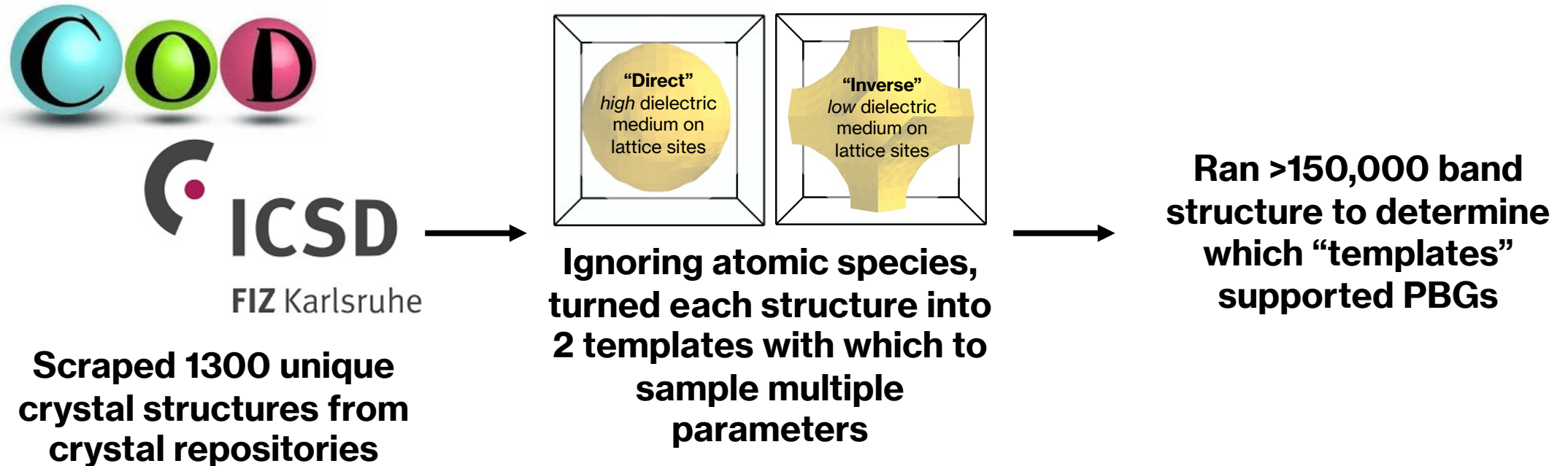
- Nanoparticles that stabilized diamond in self-assembly can **transition to lower-symmetry derivatives** at high pressure
- Small distortions in diamond **did not destroy the photonic band gap**

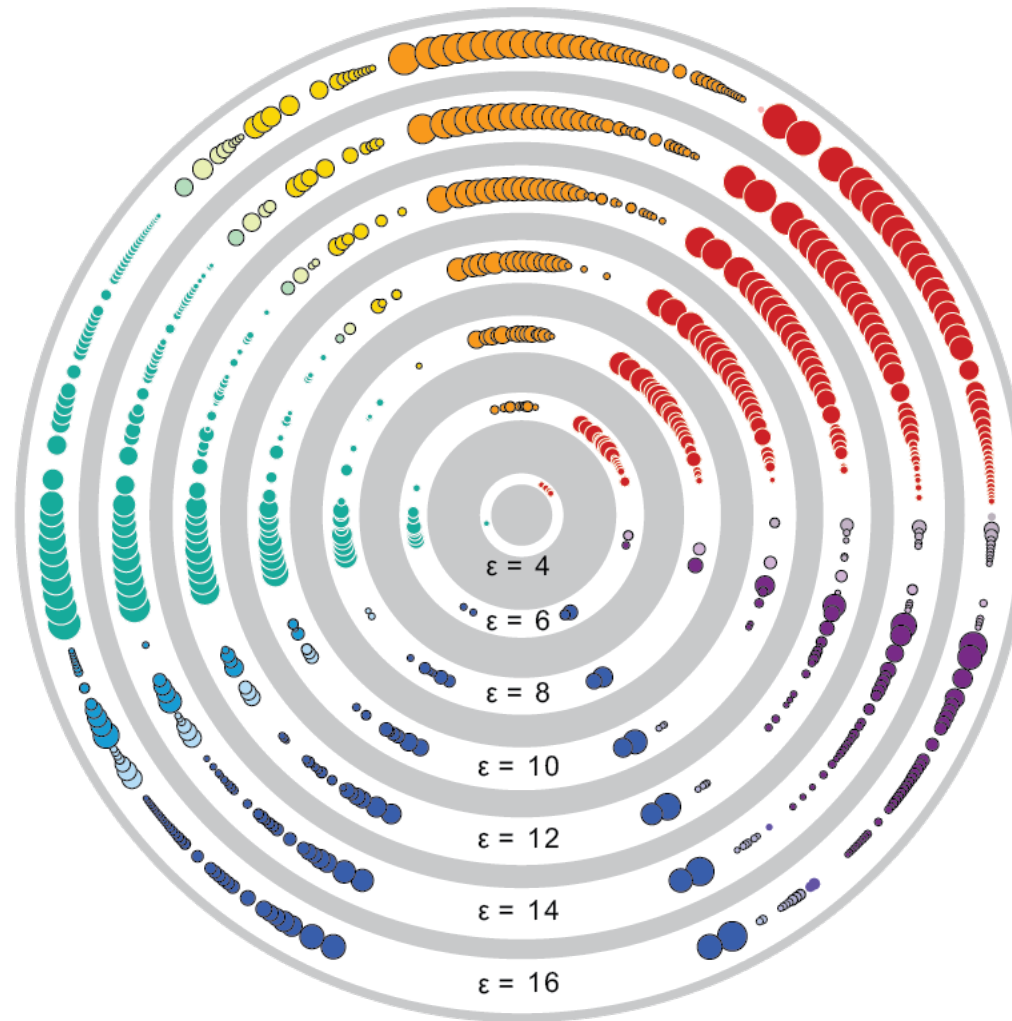
At time of presentation, this manuscript was not yet published, please see [rosecersonsky.com](http://rosecersonsky.com) for recent publications.



...Small distortions in diamond **did not destroy the photonic band gap...**  
...minimal effect on the photonic band structure...

**what is the span of crystallographic structures capable of supporting a photonic band gap?**





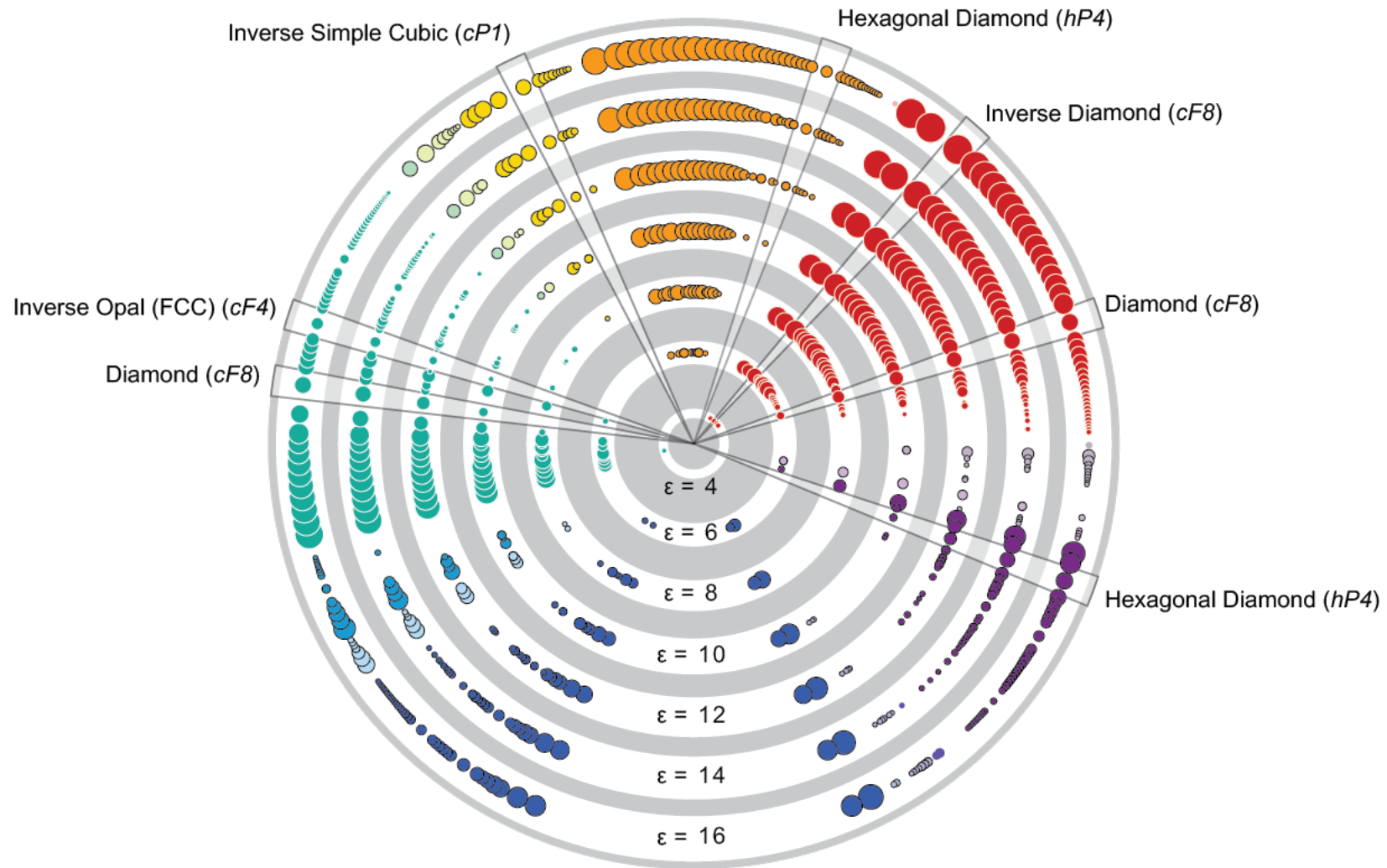
## 351 Photonic “Templates”

474 Unique Gaps

Database of Photonic Crystals:  
<https://glotzerlab.engin.umich.edu/photonics/index.html>

Appendix of Band Structures:  
<https://deepblue.lib.umich.edu/handle/2027.42/153520>

Each circle represents the maximum gap (circle size) found for a given template (radius), dielectric contrast (ring), and band location (color).

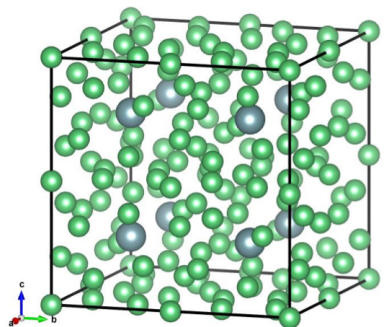
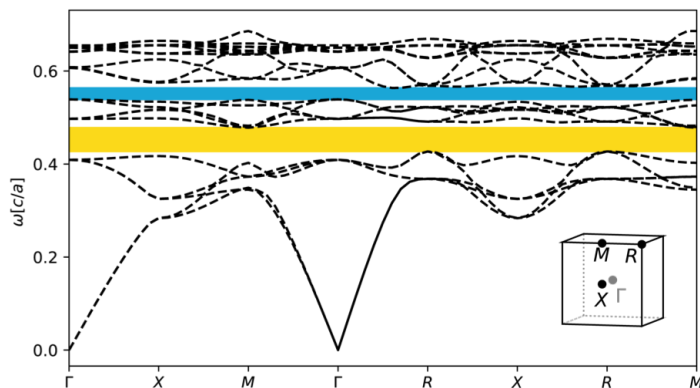


Each circle represents the maximum gap (circle size) found for a given template (radius), dielectric contrast (ring), and band location (color).



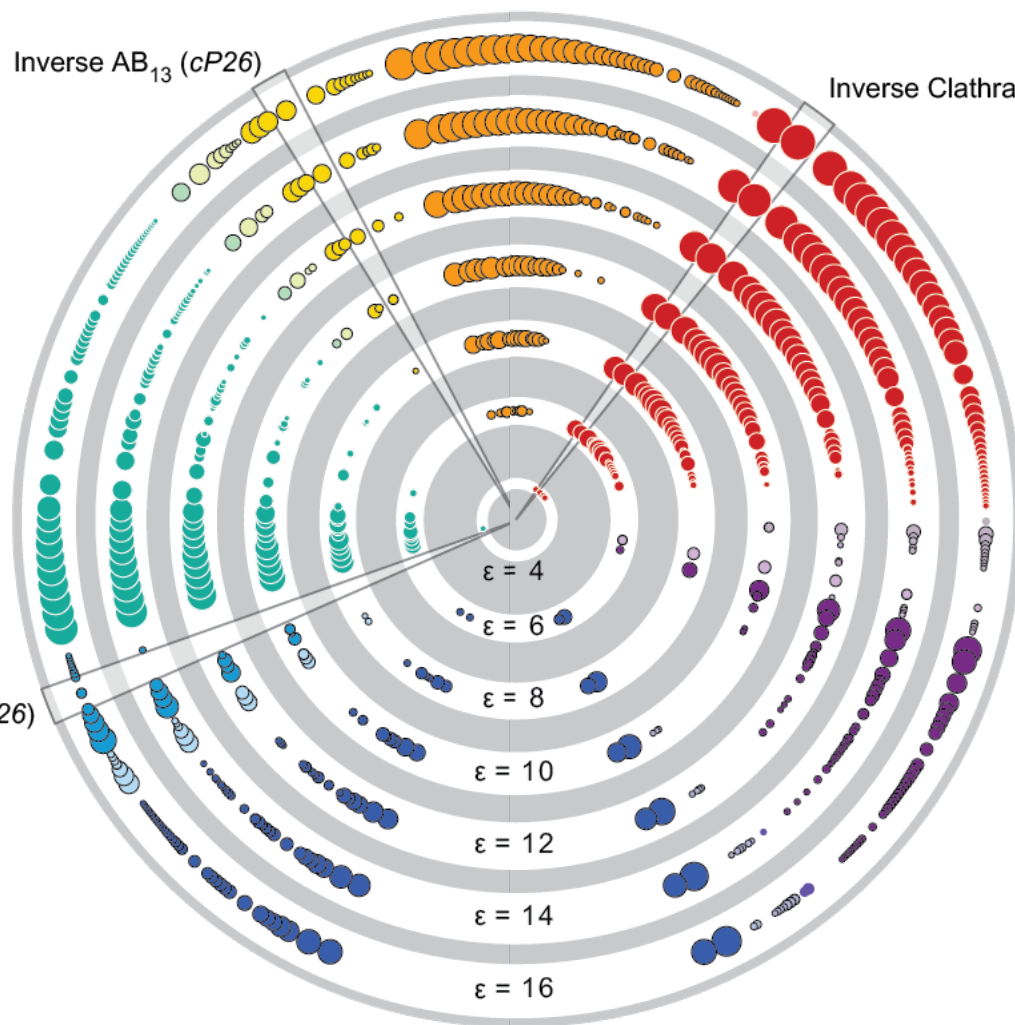
## Inverse AB<sub>13</sub>

Maximum Gaps: 13.3% (Gaps 5-6),  
4.78% (Gaps 10-11)



Inverse AB<sub>13</sub> (cP26)

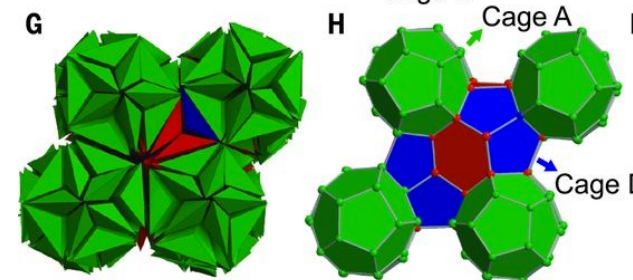
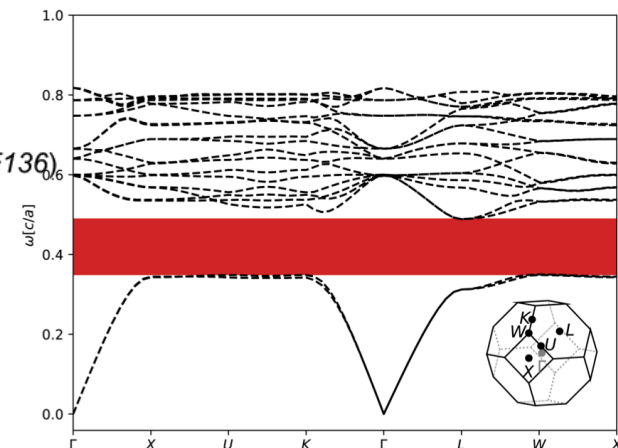
Inverse AB<sub>13</sub> (cP26)



Inverse Clathrate-II (cF136)

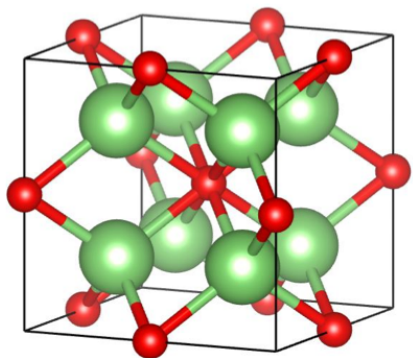
## Inverse Clathrate-II

Maximum Gap: 33.9%

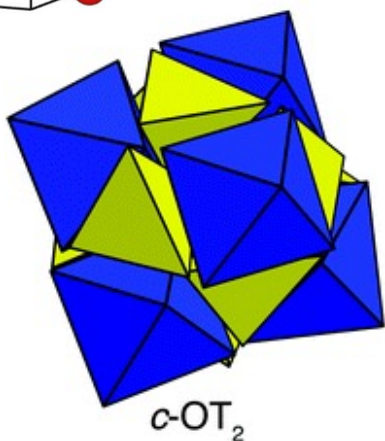


Clathrate colloidal crystals.  
Lin, H., Lee, S., Sun, L., Spellings, M.,  
Engel, M., Glotzer, S. C., & Mirkin, C. A.  
Science, 355(6328), 931-935.

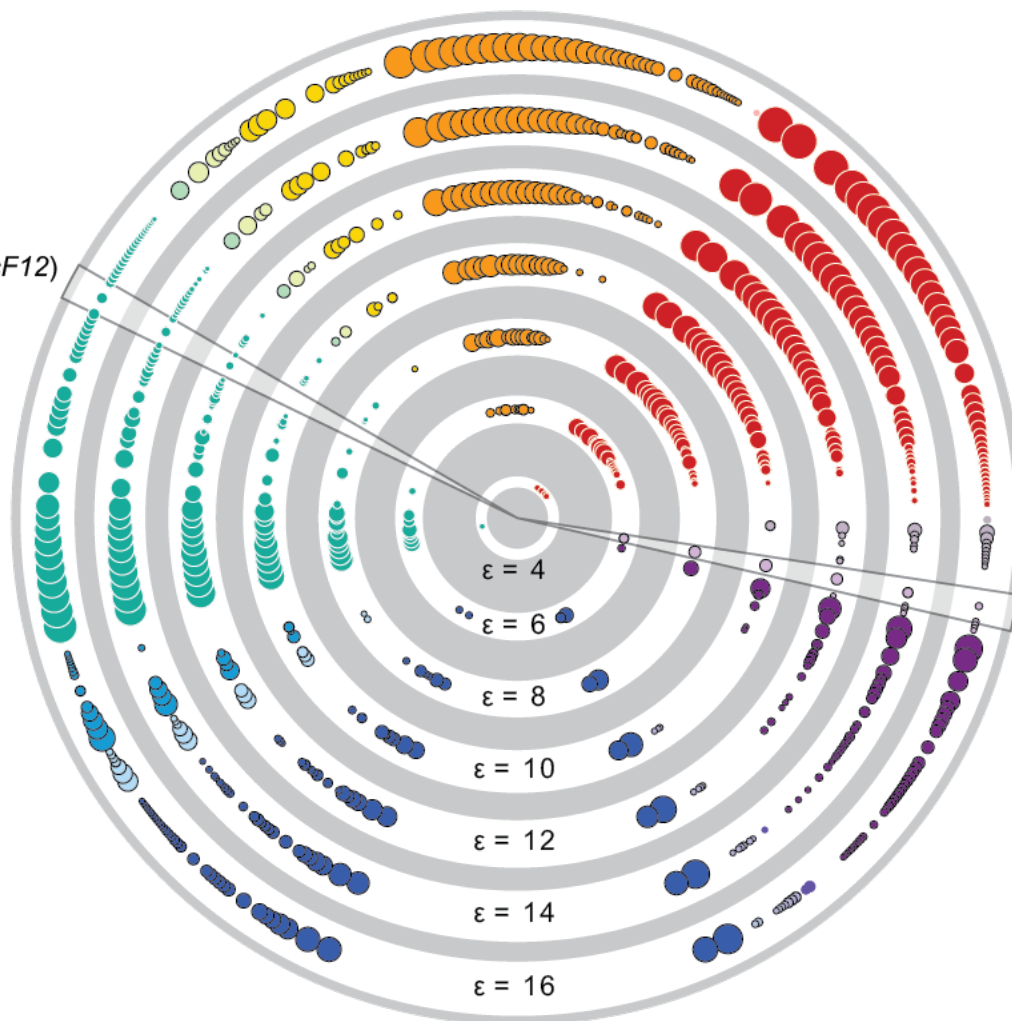
Each circle represents the maximum gap (circle size) found for a given template (radius), dielectric contrast (ring), and band location (color).



Lithium Oxide (cF12)

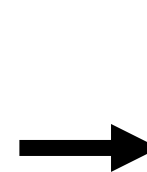


Self-assembly of a space-tessellating structure in the binary system of hard tetrahedra and octahedra.  
 Cadotte, Andrew T., et al.  
 Soft matter 12.34 (2016): 7073-7078.



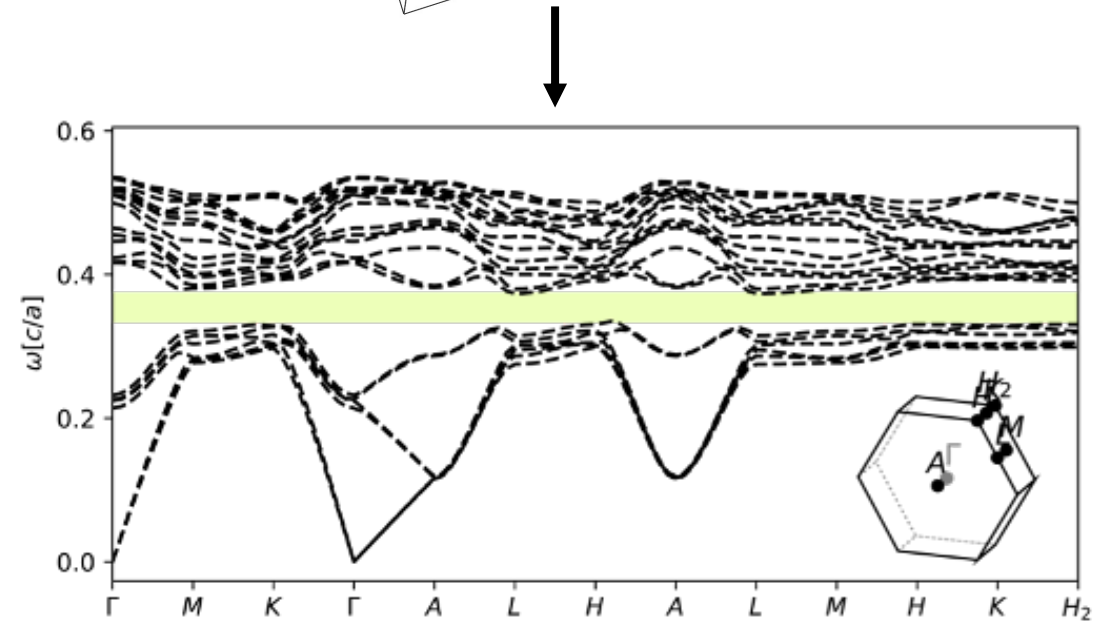
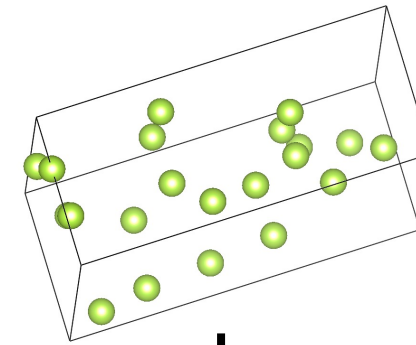
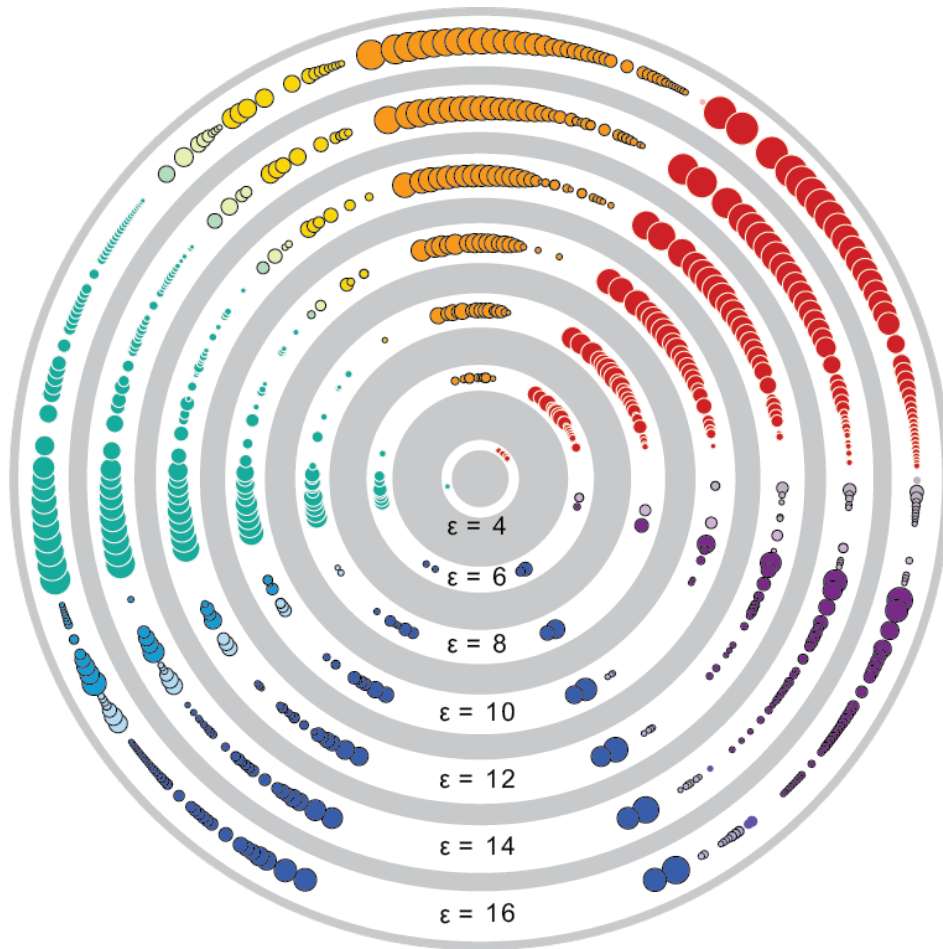
The lithium-oxide structure (a.k.a. Fluorite, c-OT<sub>2</sub>, and F-RD) exhibits photonic anomalies, including a band gap that is largest at lower dielectric contrast.

Lithium Oxide (cF12)



Each circle represents the maximum gap (circle size) found for a given template (radius), dielectric contrast (ring), and band location (color).

# How else can we use this large dataset?



Database of Photonic Crystals:  
<https://glotzerlab.engin.umich.edu/photonics/index.html>

November 7, 2021



## Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces

Jörg Behler and  
Department of Chemistry and Applied Biosciences, ETH Zurich  
(Received 27 September 2006)

The accurate description of chemical processes by methods like density-functional theory (DFT), in this Letter we introduce a new kind of neural network which provides the energy and forces as a function of several orders of magnitude faster than DFT. It is silicon and compared with empirical potentials at types of periodic and nonperiodic systems.

DOI: 10.1103/PhysRevLett.98.146401

The reliability of molecular dynamics (MD) or Monte Carlo (MC) simulations depends crucially on the accuracy of the underlying potential-energy surface (PES). *Ab initio* methods based on density-functional theory [1] (DFT) provide accurate PESs for many systems, but are computationally very demanding and even on the most advanced platforms *ab initio* MD simulations are limited to tens of picoseconds and a few thousand atoms. This is the reason for the continuing popularity of empirical potential which provide fast access to energy and forces. However, the construction of reliable empirical potentials is a difficult and lengthy process which usually relies on fitting the parameters of a guessed, physically motivated simple functional form for the interaction potential. This can lead to qualitatively wrong results when used in circumstances in which the assumed functional form is not appropriate. The database used in the fitting can include experimental or theoretical data and even the forces obtained in an *ab initio* MD run [2–4].

In this Letter we present a generalized neural-network (NN) method for constructing DFT-based PESs which have *ab initio* accuracy and are capable of describing all types of bonding. The method overcomes the limitations that have so far restricted the use of NNs to low-dimensional PES [5,6]. This is achieved by combining NN precision and flexibility with a PES representation that is inspired by empirical potentials. The resulting many-body potential are a function of all atomic coordinates and can be used in systems of arbitrary size. We apply our ideas to the construction of an NN-based many-body potential for bulk silicon. Constructing an empirical potential for Si that is valid across the phase diagram has proven to be a frustrating challenge for conventional empirical potentials. Our potential works well in the solid semiconducting and in the liquid metallic phases. In addition we can reproduce the small energy differences between the different high pressure phases of crystalline Si.

Neural networks are biology-inspired algorithms that provide an accurate tool for the representation of arbitrary functions. Given a number of points in which the value of

0031-9007/07/98(14)/146401(4)

## Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons

Albert P. Bartók,  
Cavendish Laboratory, University of Cambridge, 7 J J  
Risi

Center for the Mathematics of Information, California Insti-

Gábor Csányi,  
Engineering Laboratory, University of Cambridge, 7  
(Received 1 October 2009)

We introduce a class of interatomic potential consisting of the energies and forces experience calculations. The models do not have a fixed functional potential energy landscapes. They are systematically bulk crystals, and test it by calculating properties at quite different specific cases of a general approach angular wave numbers are used to expand the atomic in of small clusters, we quantitatively show that this ex-

DOI: 10.1103/PhysRevLett.104.136403

Atomic scale modeling of materials is now routinely and widely applied, and encompasses a range of techniques from exact quantum chemical methods [1] through density functional theory (DFT) [2] and semiempirical quantum mechanics [3] to analytic interatomic potentials [4]. The associated trade-offs in accuracy and computational cost are well known. Arguably, there is a gap between models that treat electrons explicitly and those that do not. Models in the former class are in practice limited to handling a few thousand atoms, while the simple analytic interatomic potentials are limited in accuracy, regardless of how they are parameterized. The panels in the top row of Fig. 1 illustrate the typical performance of analytic potentials in bulk semiconductors. Perhaps surprisingly, potentials that are generally regarded as adequate for describing these bulk phases show significant deviation from the quantum mechanical potential energy surface. This in turn gives rise to significant errors in predicting properties such as elastic constants and phonon spectra.

In this Letter we are concerned with the problem of modeling the Born-Oppenheimer potential energy surface (PES) of a set of atoms, but without recourse to simulating the electrons explicitly. We mostly restrict our attention to modeling the bulk phases of carbon, silicon, germanium, iron, and gallium nitride, using a unified framework. Even such single-phase potentials could be useful for calculating physical properties, e.g., the thermal expansion coefficient, the phonon contribution to the thermal conductivity, the temperature dependence of the phonon modes, or as part of hybrid schemes [5].

The first key insight is that this is actually practicable: the reason that interatomic potentials are so all-fussful is that the PES is a relatively smooth function of the nuclear coordinates. Improving potential modeling is difficult not

0031-9007/10/104(13)/136403(4)

## On representing chemical environments

Albert P. Bartók,<sup>1,2</sup> Risi K  
<sup>1</sup>Department of Engineering, University of Cambridge, 7  
<sup>2</sup>Department of Computer Science, University of Chicago  
(Received 12 December 2012; published 20 February 2013)

We review some recently published methods to represent their relative merits in terms of their faithfulness and properties that such representations (sometimes called to moving the atoms and invariance to the basis sym permutation of atoms of the same species. We demon quite different are specific cases of a general approach angular wave numbers are used to expand the atomic in of small clusters, we quantitatively show that this ex-

DOI: 10.1103/PhysRevB.87.184115

## I. INTRODUCTION

The appropriate representation of atomic environments is a crucial ingredient of algorithms used in modern computational chemistry and condensed matter physics. For example, in structure search applications,<sup>1</sup> each configuration depends numerically on the precise initial conditions and the path of the search, so it is important to be able to identify equivalent structures and detect similarities. In molecular dynamics simulations of phase transitions,<sup>2</sup> one needs good order parameters that are capable of detecting changes in the local order around atoms. Typically, the representation is in terms of a descriptor (also called a fingerprint), a tuple of real-valued functions of the atomic positions, e.g., bond lengths, bond angles, etc. “In silico” drug discovery<sup>3</sup> and other areas of chemical informatics also rely on characterizing molecules using descriptors. When constructing interatomic potentials and fitting potential energy surfaces (PES),<sup>4–8</sup> the driving application behind this work, the functional forms depend on components of a carefully chosen representation of atomic neighborhoods.

While specifying the position of each atom in a Cartesian coordinate system provides a simple and unambiguous description of atomic configurations, it is not directly suitable for making comparisons between structures: the list of coordinates is ordered arbitrarily and two structures might be mapped to each other by a rotation, reflection, or translation so that two different lists of atomic coordinates can, in fact, represent the same or very similar structures. A good representation is *invariant* with respect to permutational, rotational, reflectional, and translational symmetries, while retaining the faithfulness of the Cartesian representation. In particular, a system of invariant descriptors  $q_1, q_2, \dots, q_M$  is said to be *complete* if it uniquely determines the atomic environment up to symmetry. It is said to be *overcomplete* if it contains spurious descriptors in the sense that a proper subset of  $\{q_1, q_2, \dots, q_M\}$  is, by itself, complete. If a representation is *complete*, then

the assembly of our previously reported GDB-11 started with a collection of graphs<sup>1</sup> considered as hydrocarbons, from which chemically relevant cases were selected by topological and ring-strain criteria and expanded to produce more molecules by introducing unsaturations and heteroatoms following valency rules.<sup>2</sup> The limiting factor in computing GDB-11 was the elimination from this initial list of 98.4% of unstable and/or chemically impossible molecules using functional-group filters. Because most of the rejected molecules contained multiple heteroatoms, we reasoned that it might be possible to accelerate the database computation using a very fast “element-ratio” filter. Analysis of databases of known compounds suggested cutoff values of  $(N+O)/C < 1.0$ ,  $N/C < 0.571$ , and  $O/C < 0.666$  (see the Supporting Information). We also eliminated fluorine because it was rarely found and never considered in our group for synthesis in virtual-screening

098-0121(2013)87(18):184115(16)

184

## 970 Million Druglike Small Molecule Universe Database

Lorenz C. Blum ar  
Department of Chemistry and Materials Science, Univers  
Received March 24, 2009; E-

One of the most important chemical issues in drug discovery is innovation, in particular at the level of small organic fragments that can provide new lead structures.<sup>1</sup> The search for novel molecules can be assisted by *in silico* methods such as enumeration of chemical space,<sup>2,3</sup> breeding of molecules by genetic algorithms,<sup>4</sup> and analysis of molecular scaffolds.<sup>5</sup> We recently proposed an exhaustive enumeration approach for small organic molecules by assembling the chemical universe database GDB-11,<sup>6,7</sup> which describes the 26.4 million structures containing up to 11 atoms of C, N, O, and F that satisfy simple chemical stability and synthetic feasibility rules. We now report GDB-13, which enumerates in a similar manner small organic molecules containing up to 13 atoms of C, N, O, S, and Cl. With 977 468 314 structures, GDB-13 is the largest freely available small molecule database to date.

**Table 1. Structure Generation Statistics for GDB-13.**

nodes <sup>1</sup>	graphs <sup>2</sup>	OS <sup>3</sup>	OS <sup>4</sup>	CPU time (h) <sup>5</sup>
1	1	1	0	0.00
2	1	3	0	0.00
3	2	12	0	0.00
4	4	43	0	0.00
5	8	155	3	0.01
6	20	594	19	0.02
7	57	2 726	315	0.05
8	198	17 151	2 438	0.23
9	706	255 642	17 056	2.08
10	2 811	1 784 626	130 865	25.36
11	12 011	12 961 686	938 704	223.49
12	53 789	99 823 343	7 240 108	1 023.79
13	250 268	785 244 451	59 027 533	36 606.45
Total	319 892	910 111 673	67 356 641	39 882.86

<sup>1</sup>Number of graph nodes considered. <sup>2</sup>Number of graphs corresponding to saturated hydrocarbons passing topological and ring-strain criteria. <sup>3</sup>Molecules obtained from the graphs by combinatorial enumeration of unsaturations and heteroatoms and satisfying chemical stability and synthetic feasibility criteria. <sup>4</sup>Molecules with a reduction of CNS-containing functional groups (see the text and Supporting Information for details). <sup>5</sup>The database was compiled in parallel on a 300-node cluster (see the Supporting Information for details).

The assembly of our previously reported GDB-11 started with a collection of graphs<sup>1</sup> considered as hydrocarbons, from which chemically relevant cases were selected by topological and ring-strain criteria and expanded to produce more molecules by introducing unsaturations and heteroatoms following valency rules.<sup>2</sup> The limiting factor in computing GDB-11 was the elimination from this initial list of 98.4% of unstable and/or chemically impossible molecules using functional-group filters. Because most of the rejected molecules contained multiple heteroatoms, we reasoned that it might be possible to accelerate the database computation using a very fast “element-ratio” filter. Analysis of databases of known compounds suggested cutoff values of  $(N+O)/C < 1.0$ ,  $N/C < 0.571$ , and  $O/C < 0.666$  (see the Supporting Information). We also eliminated fluorine because it was rarely found and never considered in our group for synthesis in virtual-screening

8732 • J. AM. CHEM. SOC. 2009, 131, 8732–8733

JA I C I S  
COMMUNICATIONS

Published on Web 06/05/2009

## SCIENCE ADVANCES | RESEARCH ARTICLE

PHYSICS  
Machine learning unifies the modeling of materials and moleculesAlbert P. Bartók,<sup>1</sup> Sandip De,<sup>2,3</sup> Carl Poelking,<sup>1</sup> Noam B. Gábor Csányi,<sup>1,7</sup> Michele Ceriotti<sup>1,2,4\*</sup>

Determining the stability of molecules and condensed phases is the understanding of chemical and materials properties and transfer based on a local description of chemical environments and Bayesian to predict atomic-scale properties. It captures the quantum mechanical constructions of silicon, predicts the stability of different classes of m active and inactive protein ligands with more than 99% reliability. framework provide new insight into the potential energy surface of

**INTRODUCTION**  
Calculating the energies of molecules and condensed-phase structures is fundamental to predicting the behavior of matter at the atom scale and a formidable challenge. Reliably assessing the relative stability of different compounds, and of different phases of the same materials requires the evaluation of the energy of a given three-dimensional (3D) assembly of atoms with an accuracy comparable with the thermal energy (~4.5 kcal/mol at room temperature), which is a small fraction of the energy of a chemical bond (up to ~230 kcal/mol for the C<sub>2</sub> molecule).

Quantum mechanics is a universal framework that can deliver this level of accuracy. By solving the Schrödinger equation, the electron structure of materials and molecules can, in principle, be computed and from it all ground-state properties and excitations follow. The prohibitive computational cost of exact solutions at the level of electron structure theory leads to the development of many approximate techniques that address different classes of systems. Coupled-cluster (CC) theory (1) for molecules and density functional theory (DFT) (2–4) for the condensed phase have been particularly successful and can typically deliver the levels of accuracy required to address a plethora of important scientific questions. The computational cost of these electronic structure methods is nevertheless still significant, limiting their routine application in practice to dozens of atoms in the case of CC and hundreds in the case of DFT.

To go further, explicit electronic structure calculations have to be avoided, and we have to predict the energy corresponding to an atom configuration directly. Although such empirical potential methods (force fields) are much less expensive, their predictions to date have been qualitative at best. Moreover, the number of distinct approaches has rapidly multiplied; in the struggle for accuracy at a low cost, generality is invariably sacrificed. Recently, machine learning (ML) approaches have started to be applied to designing interatomic potentials that interpolate electronic structure data, as opposed to using pure

<sup>1</sup>Scientific Computing Department, Science and Technology Facilities Council Rutherford Appleton Laboratory, Oxfordshire OX11 0QX, UK; <sup>2</sup>National Center for Computational Design and Discovery of Novel Materials (MARVEL), Lucerne, Switzerland; <sup>3</sup>Laboratory of Computational Science and Modeling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; <sup>4</sup>Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK; <sup>5</sup>Center for Materials Physics and Technology, U.S. Naval Research Laboratory, Washington, DC 20375, USA; <sup>6</sup>Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, UK; <sup>7</sup>Engineering Laboratory, University of Cambridge, Cambridge, UK

\*Corresponding author. Email: michele.cerotti@epfl.ch

Bartók et al., Sci. Adv. 2017;3:e1701816 13 December 2017

## SCIENCE ADVANCES | RESEARCH ARTICLE

APPLIED MATHEMATICS  
Machine learning of accurate energy-conserving molecular force fieldsStefan Chmiela,<sup>1</sup> Alexandre Tkatchenko,<sup>2,3\*</sup> Huziel E. Sauceda,<sup>3</sup> Igor Potlitsky,<sup>2</sup> Kristof T. Schütt,<sup>1</sup> Klaus-Robert Müller<sup>1,4,5,6\*</sup>

Using conservation of energy—a fundamental property of closed classical and quantum mechanical systems—we develop an efficient gradient-domain machine learning (GDML) approach to construct accurate molecular force fields using a reduced number of samples from *ab initio* molecular dynamics (AIMD) trajectories. The GDML implementation is able to reproduce global potential energy surfaces of intermediate-sized molecules with an accuracy of 0.3 kcal mol<sup>-1</sup> for energies and 1 kcal mol<sup>-1</sup> Å<sup>-1</sup> for atomic forces using only 1000 conformational geometries for training. We demonstrate this accuracy for AIMD trajectories of molecules, including benzene, toluene, naphthalene, ethanol, urea, and aspirin. The challenge of constructing conservative force fields is accomplished in our work by learning in a Hilbert space of vector-valued functions that obey the law of energy conservation. The GDML approach enables quantitative molecular dynamics simulations for molecules at a fraction of cost of explicit AIMD calculations, thereby allowing the construction of efficient force fields with the accuracy and transferability of high-level *ab initio* methods.

**INTRODUCTION**  
Within the Born-Oppenheimer (BO) approximation, predictive simulations of properties and functions of molecular systems require an accurate description of the global potential energy hypersurface  $V_{\text{BO}}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ , where  $\mathbf{r}_i$  indicates the nuclear Cartesian coordinates. Although  $V_{\text{BO}}$  could, in principle, be obtained on the fly using explicit *ab initio* calculations, more efficient approaches that can access the long time scales are required to understand relevant phenomena in large molecular systems. A plethora of classical mechanistic approximations to  $V_{\text{BO}}$  have been constructed, in which the parameters are typically fitted to a small set of *ab initio* calculations or experimental data. Unfortunately, these classical approximations may suffer from the lack of transferability and can yield accurate results only close to the conditions (geometries) they have been fitted to. Alternatively, sophisticated machine learning (ML) approaches that can accurately reproduce the global potential energy surface (PES) for elemental materials (1–9) and small molecules (10–16) have been recently developed (see Fig. 1, A and B) (17). Although potentially very promising, one particular challenge for direct ML fitting of molecular PES is the large amount of data necessary to obtain an accurate model. Often, many thousands or even millions of atomic configurations are used as training data for ML models. This results in nontransparent models, which are difficult to analyze and may break consistency (18) between energies and forces.

A fundamental property that any force field  $\mathbf{F}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$  must satisfy is the conservation of total energy, which implies that  $\mathbf{F}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = -\nabla_{\mathbf{r}_i} V(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ . Any classical mechanistic expressions for the potential energy (also denoted as classical force field) or analytically derivable ML approaches trained on energies satisfy energy conservation by construction. However, even if conserv-

tion of energy is satisfied implicitly within an approximation, this does not imply that the model will be able to accurately follow the trajectory of the true *ab initio* potential, which was used to fit the force field. In particular, small energy/force inconsistencies between the force field model and *ab initio* calculations can lead to unforeseen artifacts in the PES topology, such as spurious critical points that can give rise to incorrect molecular dynamics (MD) trajectories. Another fundamental problem is that classical and ML force fields focusing on energy as the main observable have to assume atomic energy additivity—an approximation that is hard to justify from quantum mechanics.

Here, we present a robust solution to these challenges by constructing an explicitly conservative ML force field, which uses exclusively atom gradient information in lieu of atomic (or total) energies. In this manner, with any number of data points, the proposed model fulfills energy conservation by construction. Obviously, the developed ML force field can be coupled to a heat bath, making the full system (molecule and bath) non-energy-conserving.

We remark that atomic forces are true quantum-mechanical observables within the BO approximation by virtue of the Hellmann-Feynman theorem. The energy of a molecular system is recovered by analytical integration of the force-field kernel (see Fig. 1C). We demonstrate that our gradient-domain machine learning (GDML) approach is able to accurately reproduce global PESs of intermediate-sized molecules within 0.3 kcal mol<sup>-1</sup> for energies and 1 kcal mol<sup>-1</sup> Å<sup>-1</sup> for atomic forces relative to the reference data. This accuracy is achieved when using less than 1000 training geometries to construct the GDML model and using energy conservation to avoid overfitting and artifacts. Hence, the GDML approach paves the way for efficient and precise MD simulations with PESs that are obtained with arbitrary high-level quantum-chemical approaches. We demonstrate the accuracy of GDML by computing AIMD-quality thermodynamic observables using path-integral MD (PIMD) for eight organic molecules with up to 21 atoms and four chemical elements. Although we use density functional theory (DFT) calculations as reference in this development work, it is possible to use any higher-level quantum-chemical reference data. With state-of-the-art quantum chemistry codes running on current high-performance computers, it is possible to generate accurate reference data for molecules with a few dozen atoms. Here, we focus on intramolec-

Chmiela et al., Sci. Adv. 2017;3:e1603015 5 May 2017

1 of 6

# EPFL



**COSMO**  
COMPUTATIONAL SCIENCE AND MODELLING

Machine Learning Science and Technology

PAPER

RECEIVED: 18 December 2020  
REVISED: 23 April 2021  
ACCEPTED FOR PUBLICATION: 6 May 2021  
PUBLISHED: 14 July 2021

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 license.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

DOI: 10.21105/joss.01903

Software

- Review it
- Repository it
- Archive it

Editor: Christopher Okbin

Submitted: 12 Feb  
Published: 01 July

License: Authors of papers re copyright and release under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

**JOSS**  
The Journal of Open Source Software

Machine Learning Science and Technology

PAPER

OPEN ACCESS

### Improving sample and feature selection with principal covariates regression

Rose K Cersosky<sup>1\*</sup>, Benjamin A Helfrecht<sup>2</sup>, Edgar A Engel<sup>2</sup>, Sergei Klavinec<sup>2</sup> and Michele Ceriotti<sup>1\*</sup>

<sup>1</sup> Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland  
<sup>2</sup> TCM Group, Cavendish Laboratory, University of Cambridge, JJ, Thomson Avenue, Cambridge CB3 0HF, United Kingdom

\* Authors to whom any correspondence should be addressed.  
E-mail: rose.cersosky@epfl.ch and michele.ceriotti@epfl.ch

Keywords: machine learning, feature selection, sample selection, farthest point sampling, materials science, physical chemistry, semi-supervised learning

Supplementary material for this article is available online

Abstract  
Selecting the most relevant features and samples out of a large set of candidates is a task that occurs very often in the context of automated data analysis, where it improves the computational performance and often the transferability of a model. Here we focus on two popular subselection schemes applied to this end: CUR decomposition, derived from a low-rank approximation of the feature matrix, and farthest point sampling (FPS), which relies on the iterative identification of the most diverse samples and discriminating features. We modify these unsupervised approaches, incorporating a supervised component following the same spirit as the principal covariates (PCov) regression method. We show how this results in selections that perform better in supervised tasks, demonstrating with models of increasing complexity, from ridge regression to kernel ridge regression and finally feed-forward neural networks. We also present adjustments to minimise the impact of any subselection when performing unsupervised tasks. We demonstrate the significant improvements associated with PCov-CUR and PCov-FPS selections for applications to chemistry and materials science, typically reducing by a factor of two the number of features and samples required to achieve a given level of regression accuracy.

203: Innovations in Methods of Data Science  
Monday, November 8, 2021, 4:30 PM - 4:45 PM  
Marriott Copley Place - Salon H/I

At time of presentation, this manuscript was not yet published, please see [rosecersosky.com](https://rosecersosky.com) for recent publications.

At time of presentation, this work was not yet published, please see [rosepersonskey.com](http://rosepersonskey.com) for recent publications.



# The Search for Novel Mesoscale Materials

**Rose K. Cersonsky**

Laboratory of Computational Science and Modeling (COSMO)  
École Polytechnique Fédérale de Lausanne (EPFL)  
Lausanne, Switzerland

At time of presentation, this work was not yet published,  
please see [rosecersonsky.com](http://rosecersonsky.com) for recent publications.

**Come see my last talk!**  
**203e - Improving Data  
Sub-Selection for  
Supervised Tasks with  
Principal Covariates  
Regression**

Monday, November 8,  
2021  
4:30 PM - 4:45 PM EDT  
Marriott Copley Place -  
Salon H/I



[rosecersonsky.com](http://rosecersonsky.com)