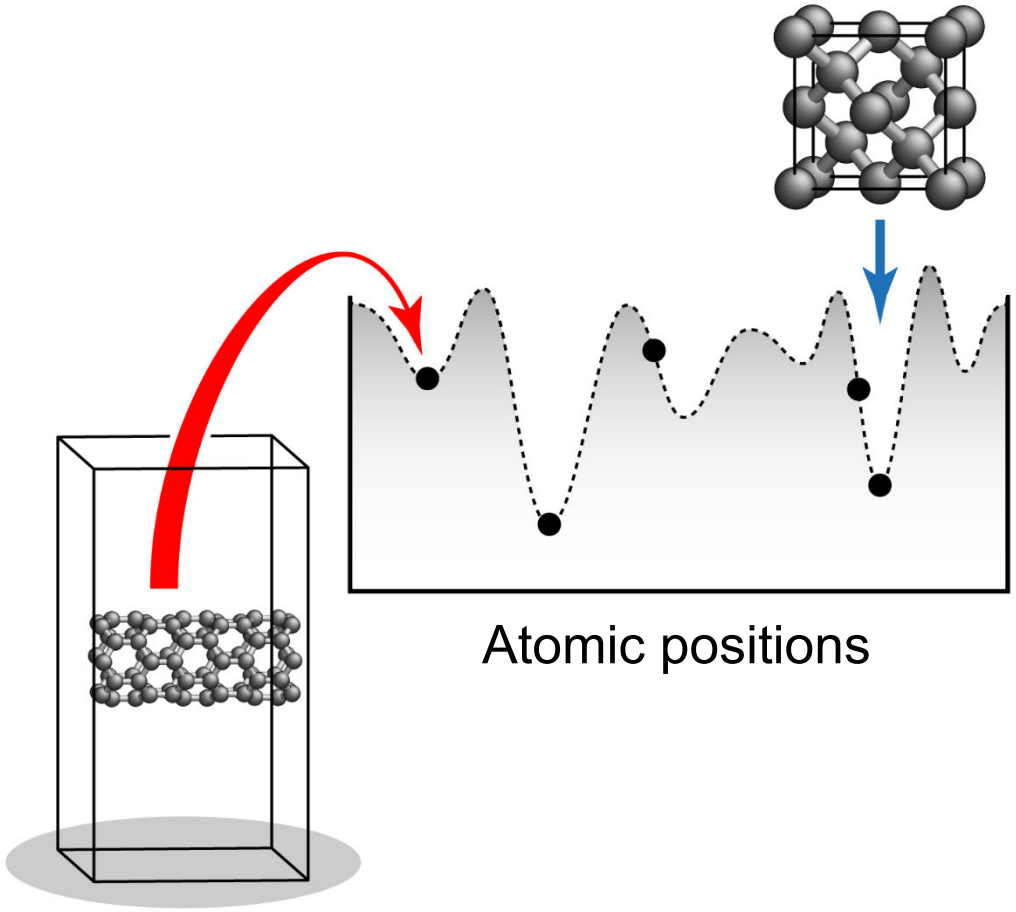
Gaussian process regression for atomistic materials modelling

Volker Deringer with tutorials by Joe Morrow

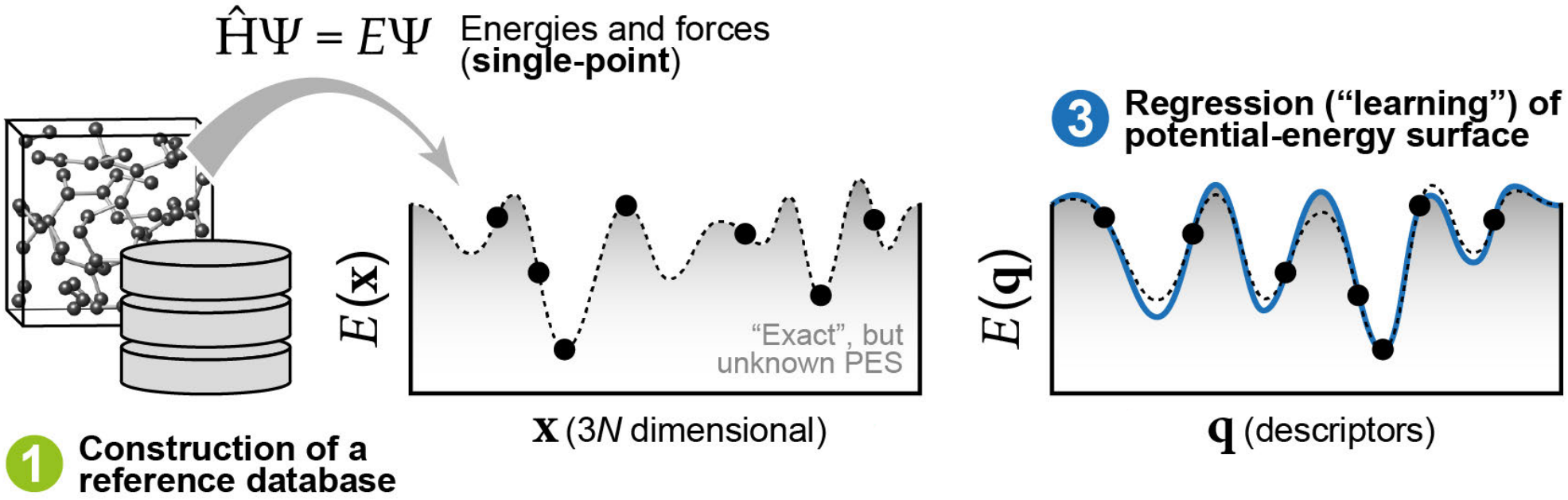
http://deringer.chem.ox.ac.uk

Machine-learning-driven materials modelling



Machine-learning-driven materials modelling

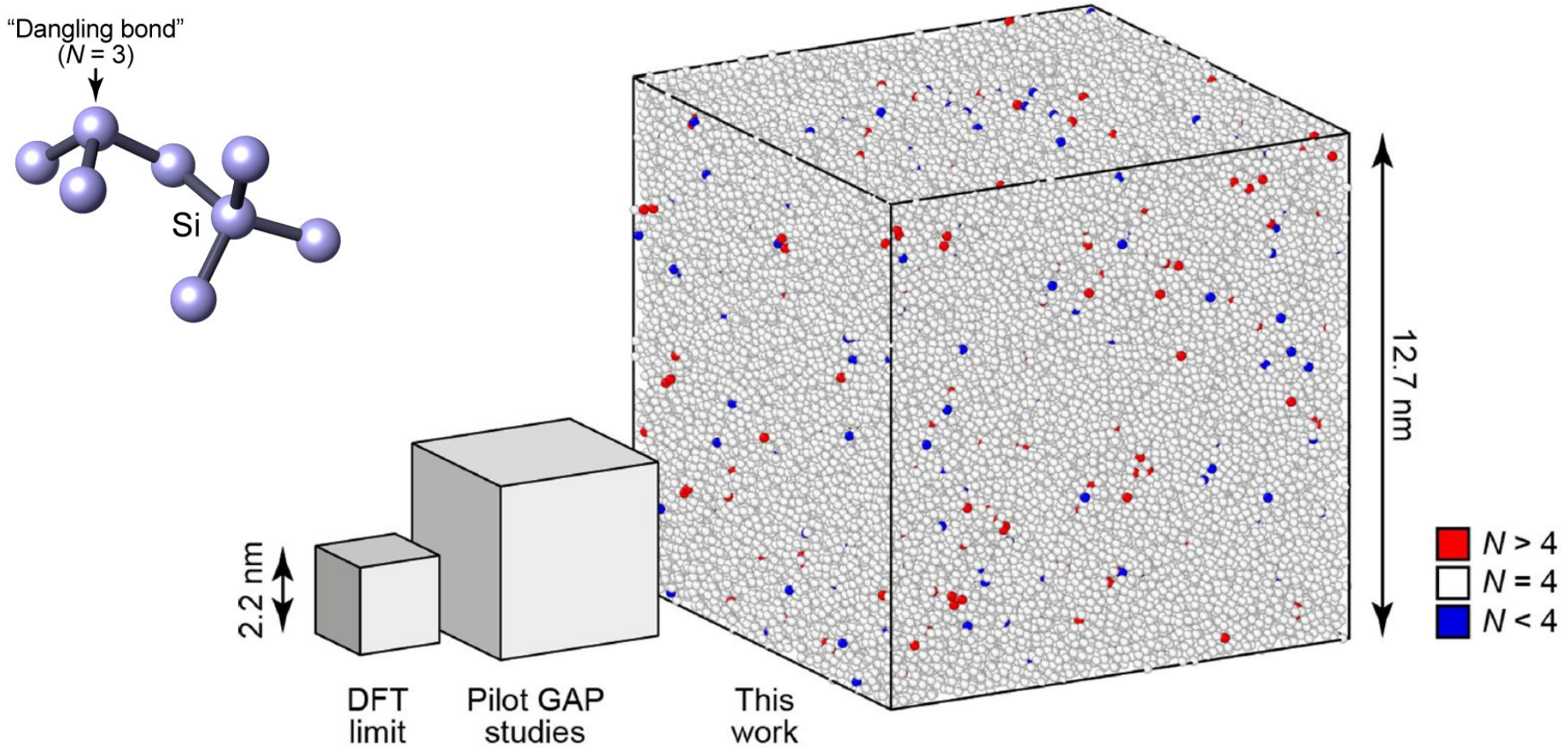
ML potentials approximate the quantum-mechanical potential-energy surface, using three main **ingredients**:



Adv. Mater. **2019**, *31*, 1902765

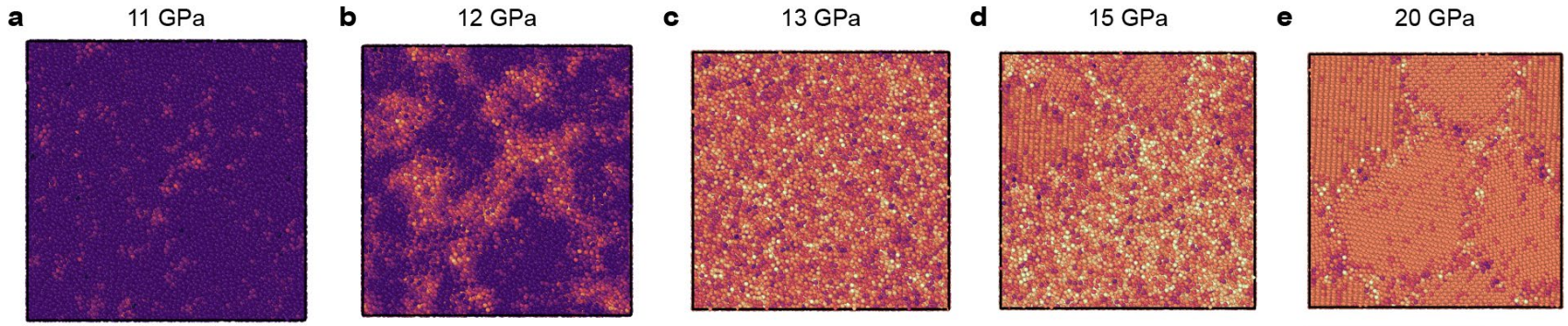
A “textbook case” for random networks: **Amorphous silicon**

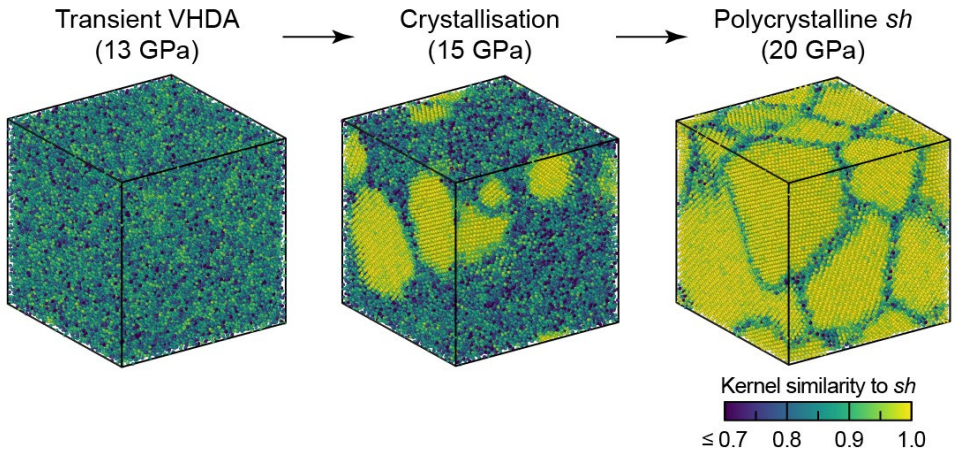
Silicon “GAP-18”: Bartók et al., *Phys. Rev. X* **2018**, *8*, 041048
Validation for *a*-Si: VLD et al., *J. Phys. Chem. Lett.* **2018**, *9*, 2879

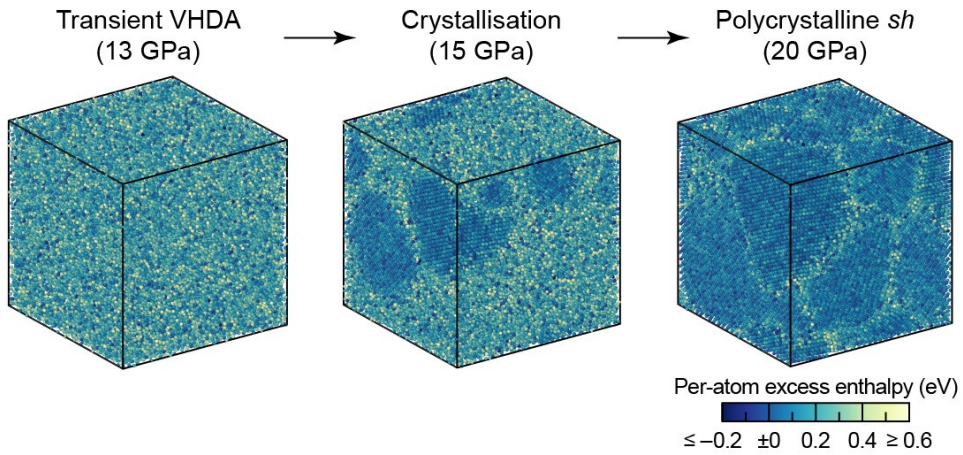


VLD, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, *Nature* **2021**, *589*, 59

Origins of structural transitions in disordered silicon

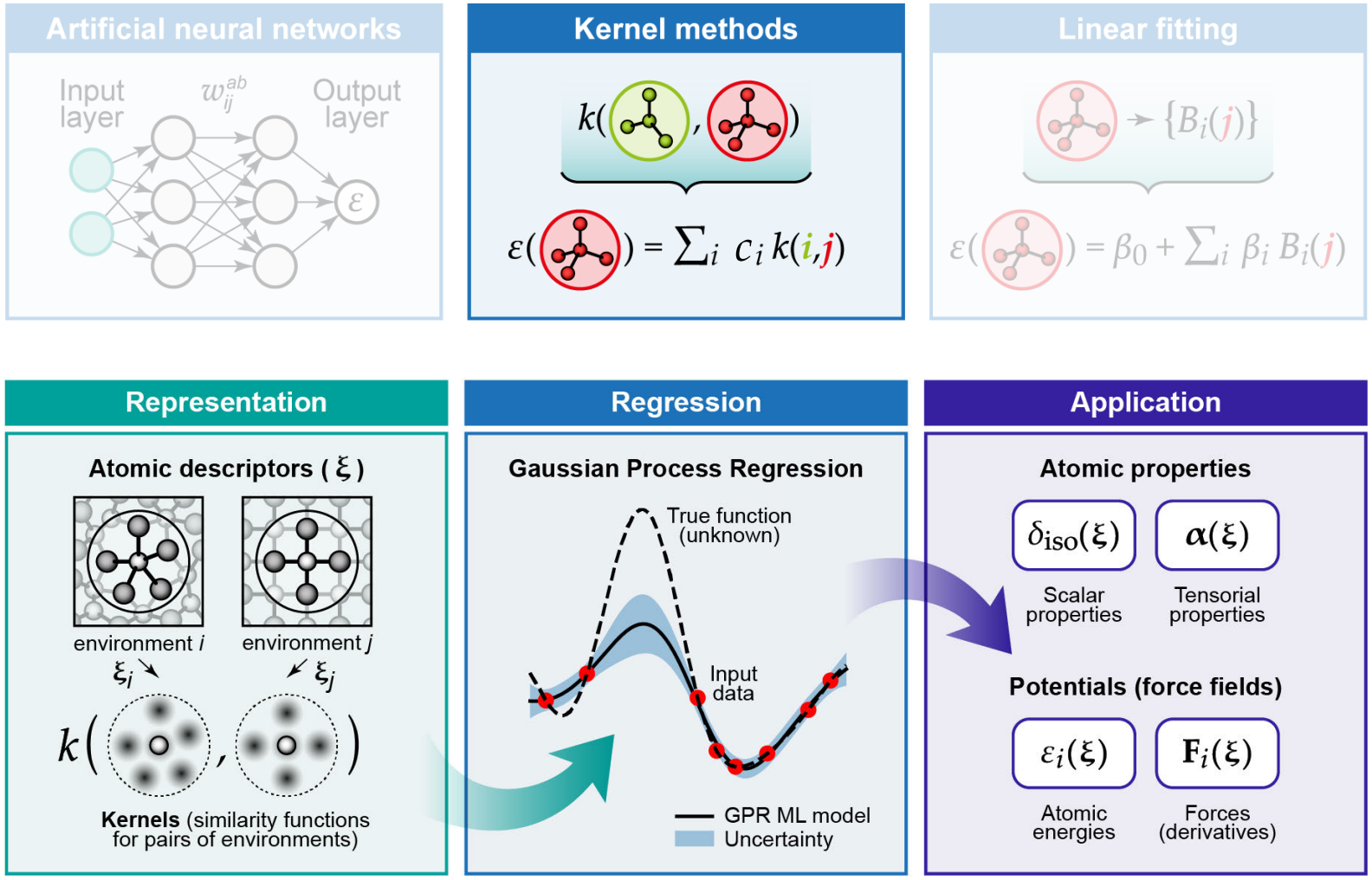


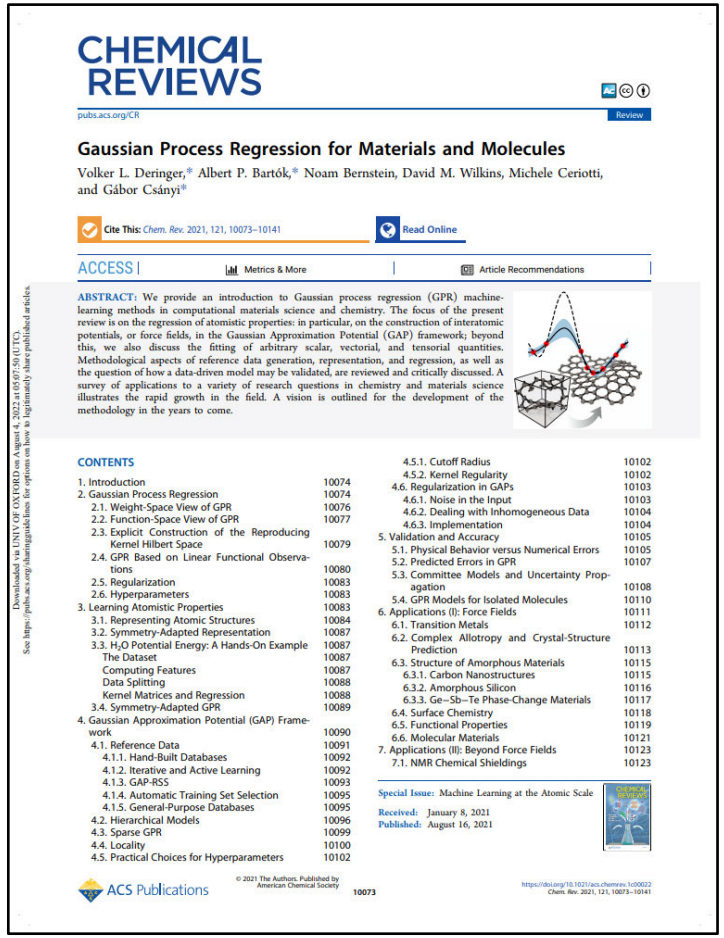




Nature **2021**, *589*, 59

Gaussian process regression (GPR) for materials




CHEMICAL REVIEWS

Gaussian Process Regression for Materials and Molecules

Volker L. Deringer,* Albert P. Bartók,* Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi*

Cite This: *Chem. Rev.* 2021, 121, 10073−10141 | Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: We provide an introduction to Gaussian process regression (GPR) machine-learning methods in computational materials science and chemistry. The focus of the present review is on the regression of atomistic properties: in particular, on the construction of interatomic potentials, or force fields, in the Gaussian Approximation Potential (GAP) framework; beyond this, we also discuss the fitting of arbitrary scalar, vectorial, and tensorial quantities. Methodological aspects of reference data generation, representation, and regression, as well as the question of how a data-driven model may be validated, are reviewed and critically discussed. A survey of applications to a variety of research questions in chemistry and materials science illustrates the rapid growth in the field. A vision is outlined for the development of the methodology in the years to come.

CONTENTS

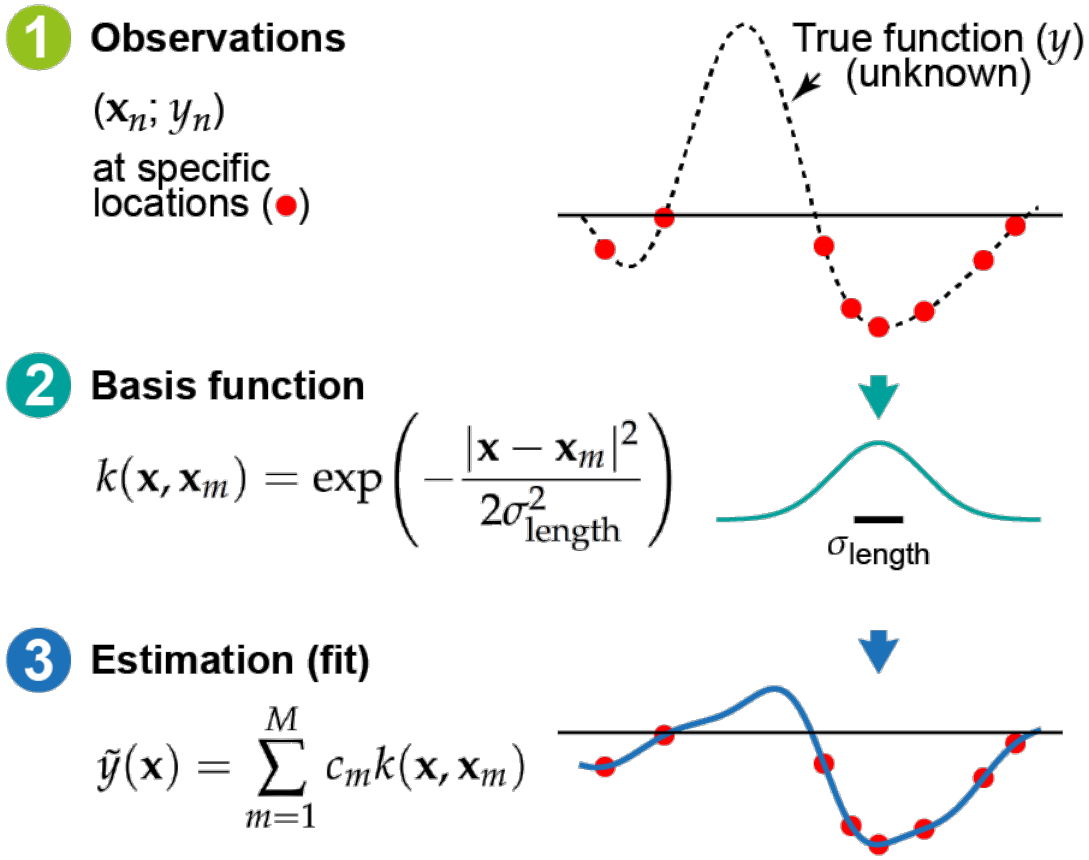
Special Issue: Machine Learning at the Atomic Scale
Received: January 8, 2021
Published: August 16, 2021

ACS Publications

https://doi.org/10.1021/acs.chemrev.1c00022
Chem. Rev. 2021, 121, 10073−10141

Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

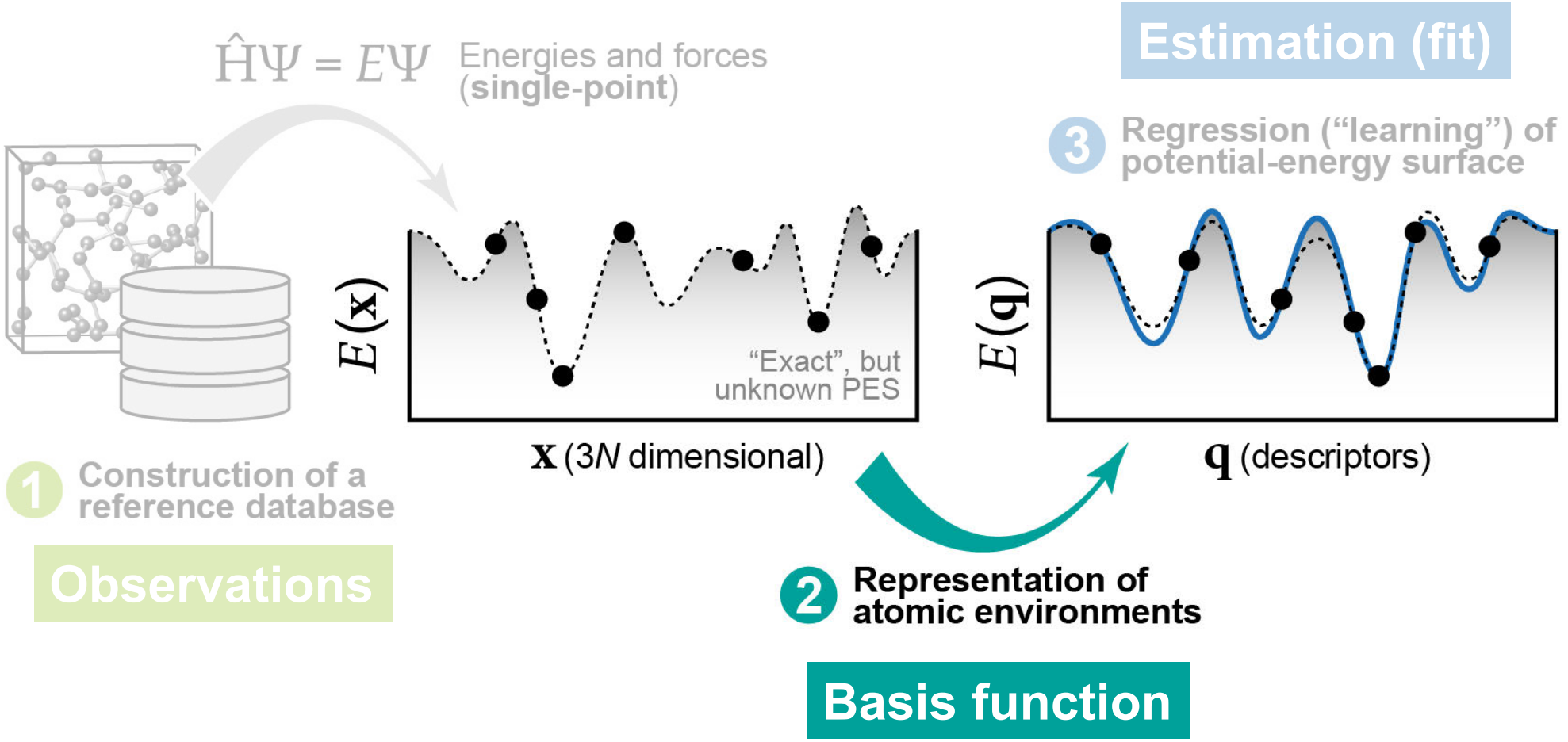
Gaussian process regression (GPR)



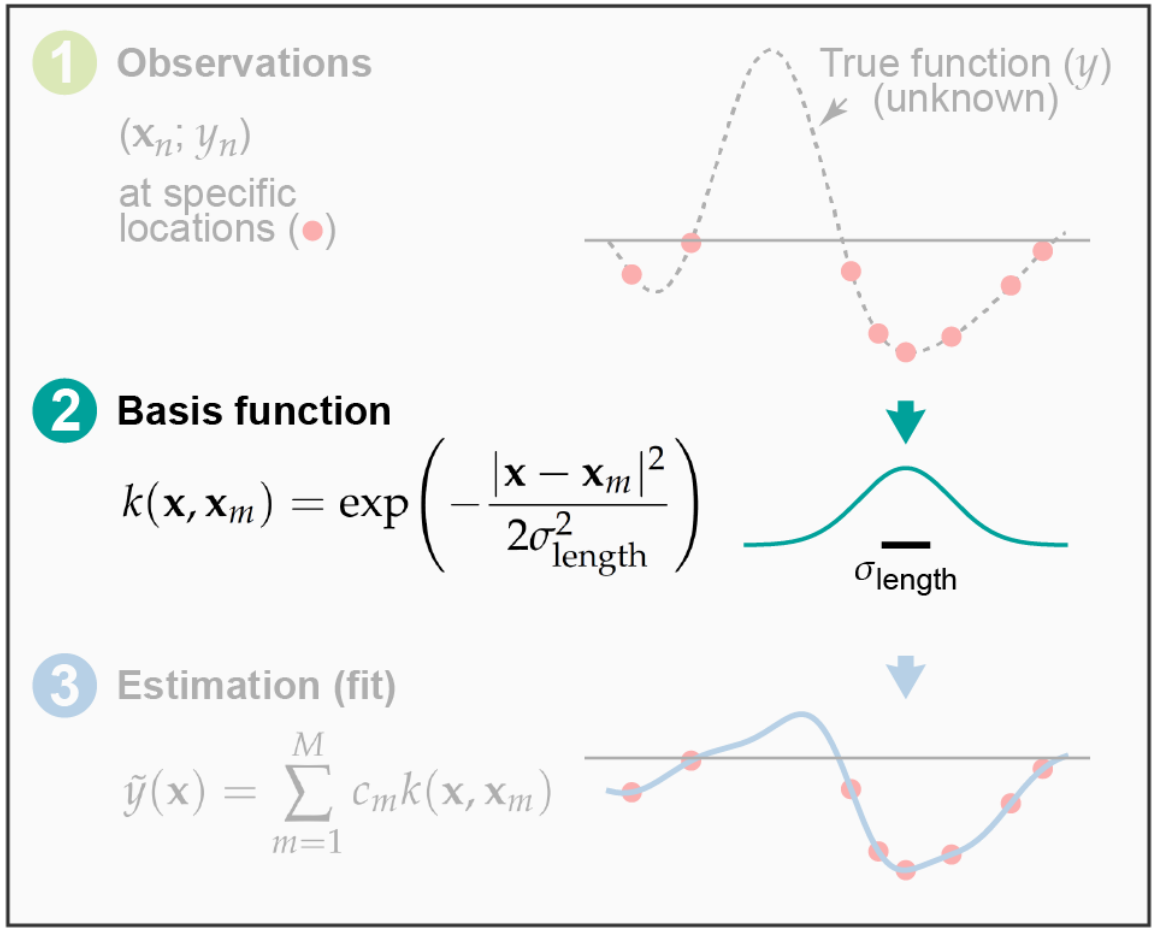
- **Observations**, using the "ground truth", make up the training data
- In atomistic ML: **x** encodes atomic environment, *y* can be any per-atom property
- **Basis functions** are built using a kernel similarity measure, *k*
- In atomistic ML: from simple pair distance to more complex forms
- **Estimation** means predicting *y* at a new, unknown location **x**
- To do this, we need *k* and pre-determined fitting coefficients, *c*

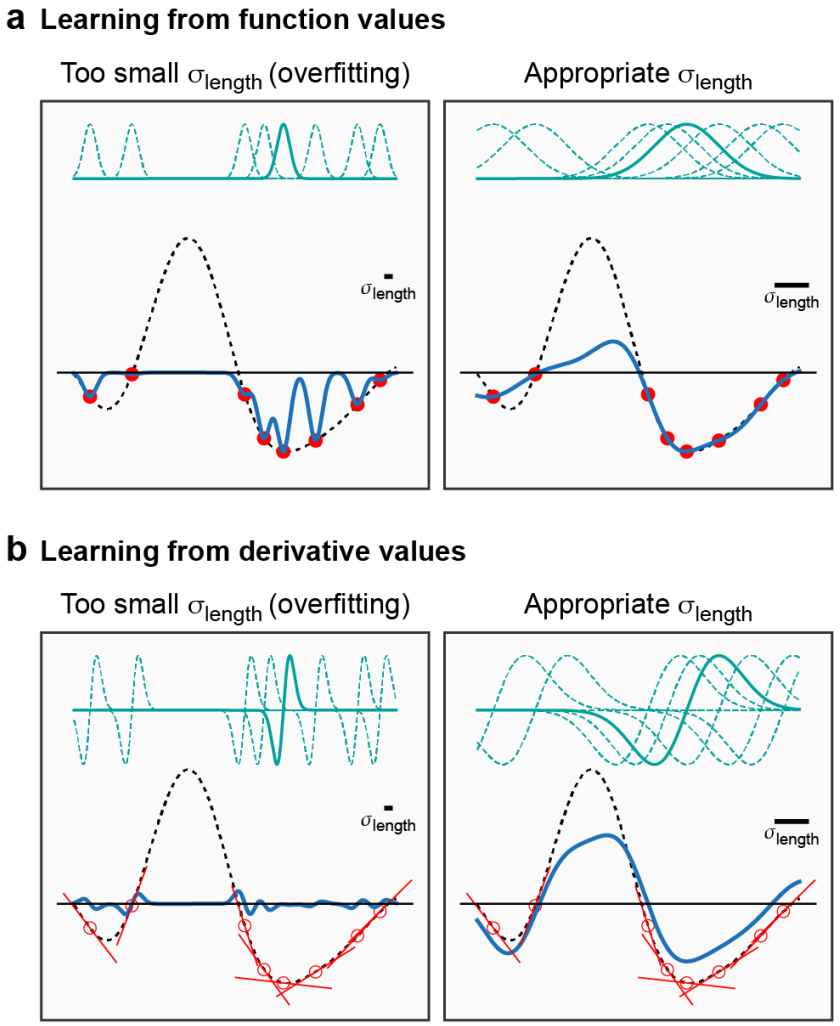
Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR) for materials



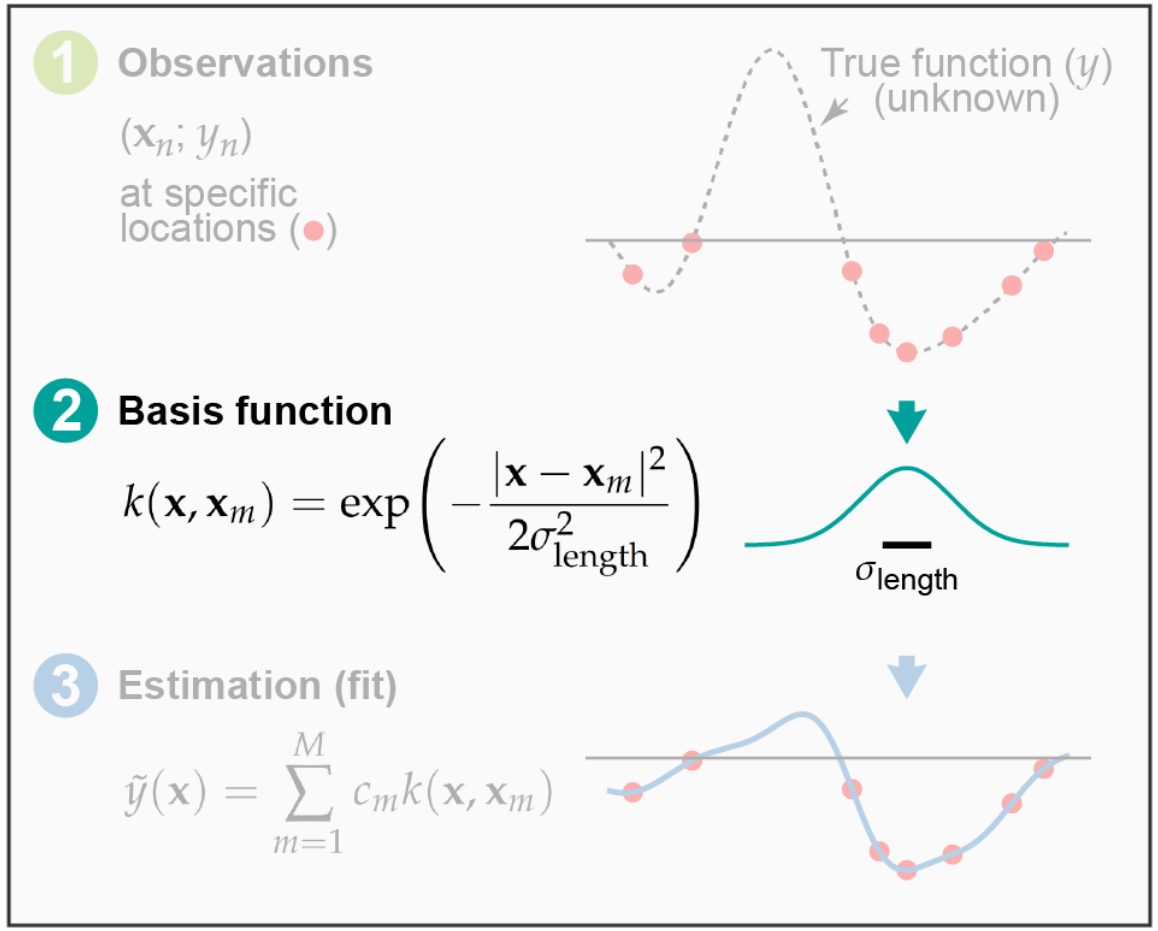
Gaussian process regression (GPR)





Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

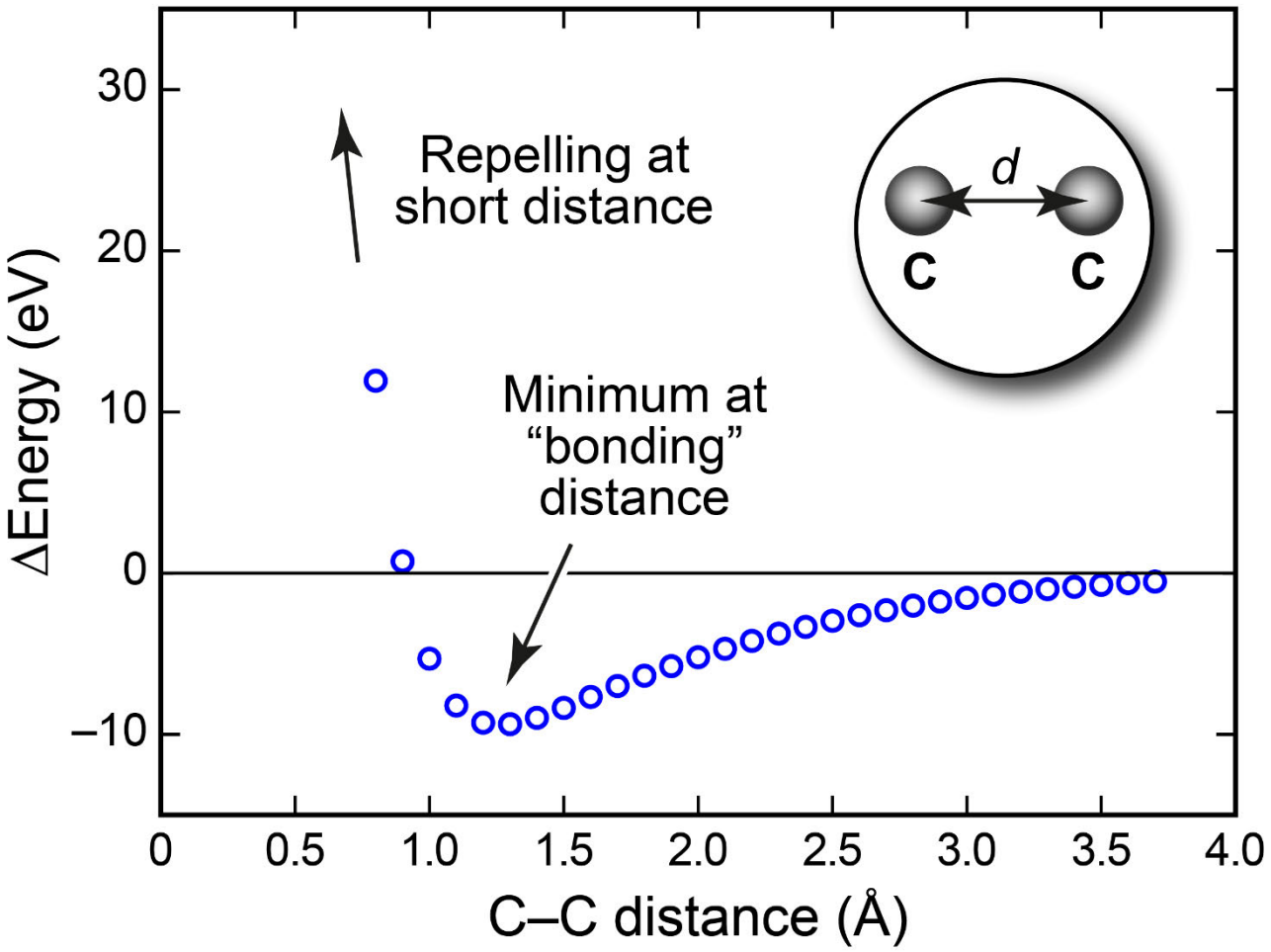
Gaussian process regression (GPR)



- **Observations**, using the "ground truth", make up the training data
- In atomistic ML: **x** encodes atomic environment, *y* can be any per-atom property
- **Basis functions** are built using a kernel similarity measure, *k*
- In atomistic ML: from simple pair distance to more complex forms
- **Estimation** means predicting *y* at a new, unknown location **x**
- To do this, we need *k* and pre-determined fitting coefficients, *c*

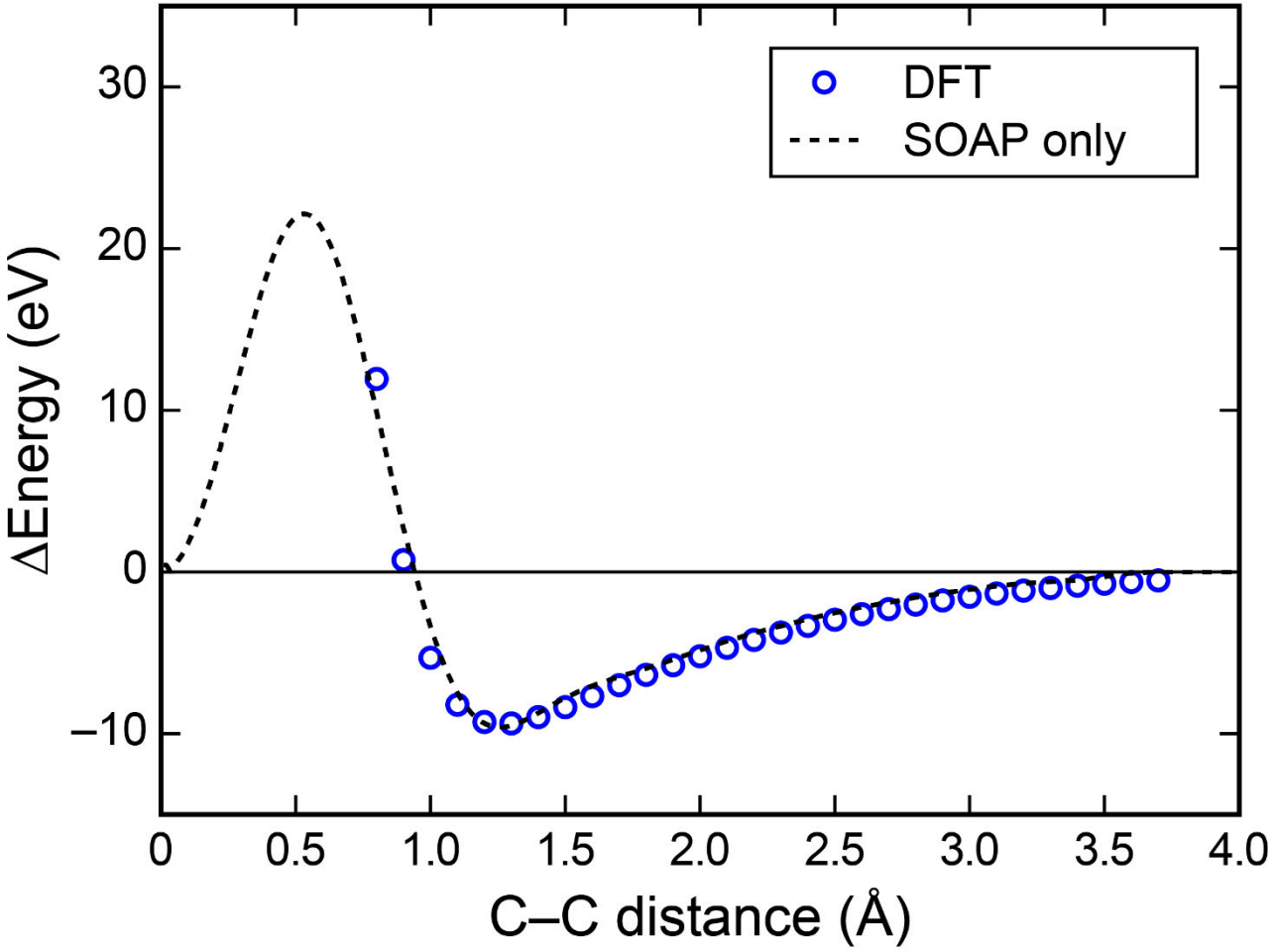
Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Representations (*descriptors*) in practice

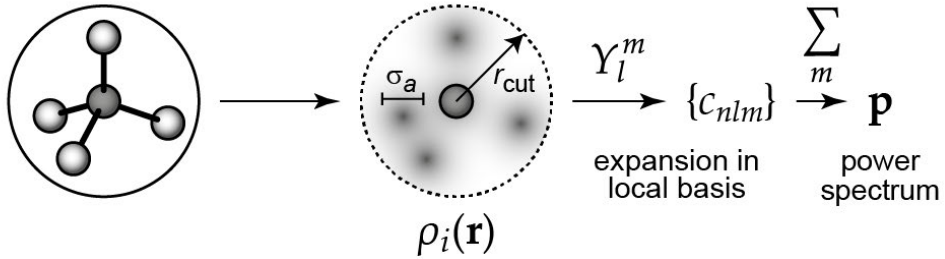


VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

Representations (*descriptors*) in practice



Use Smooth Overlap of Atomic Positions (**SOAP**) only to encode atomic structure

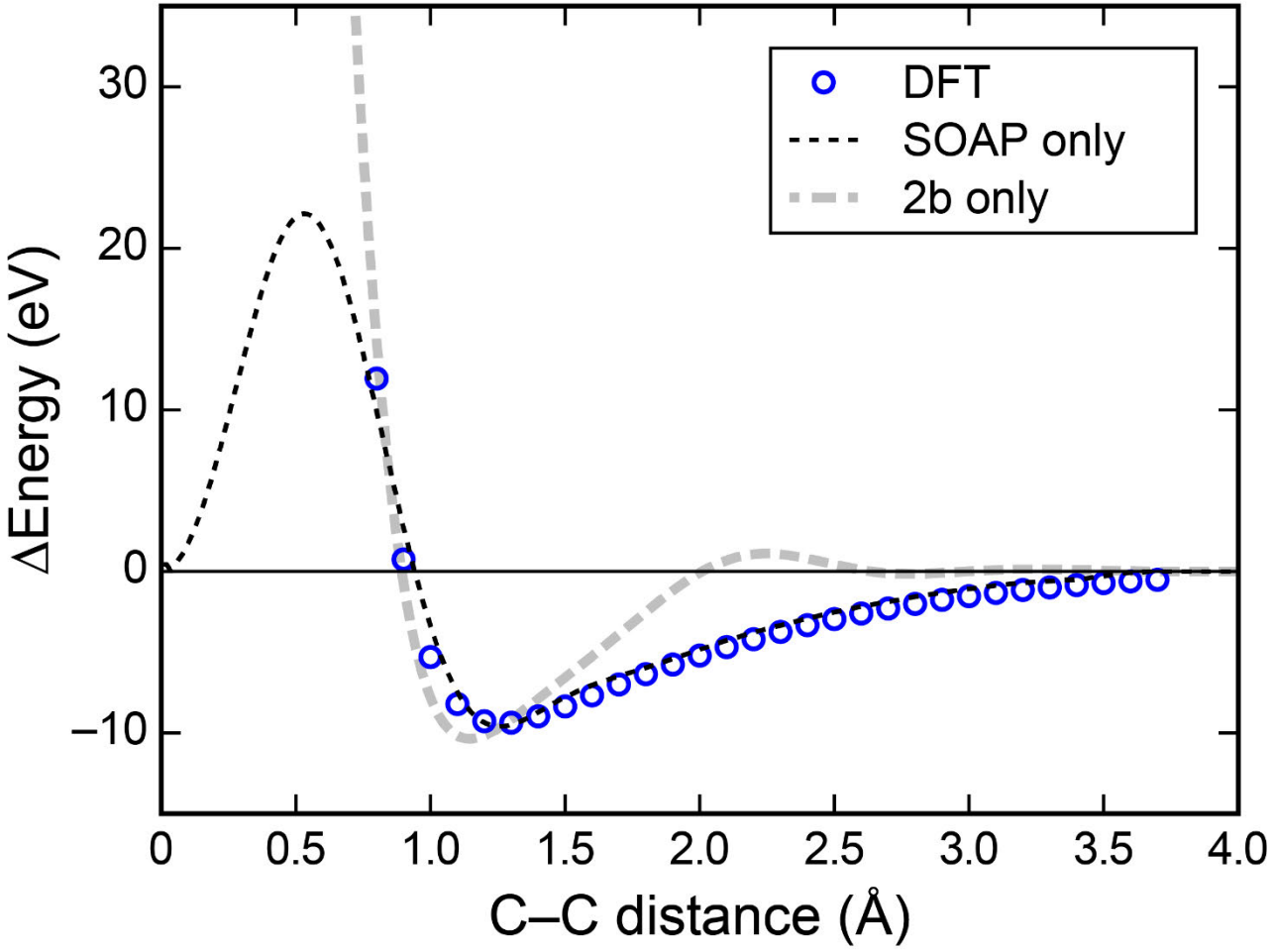


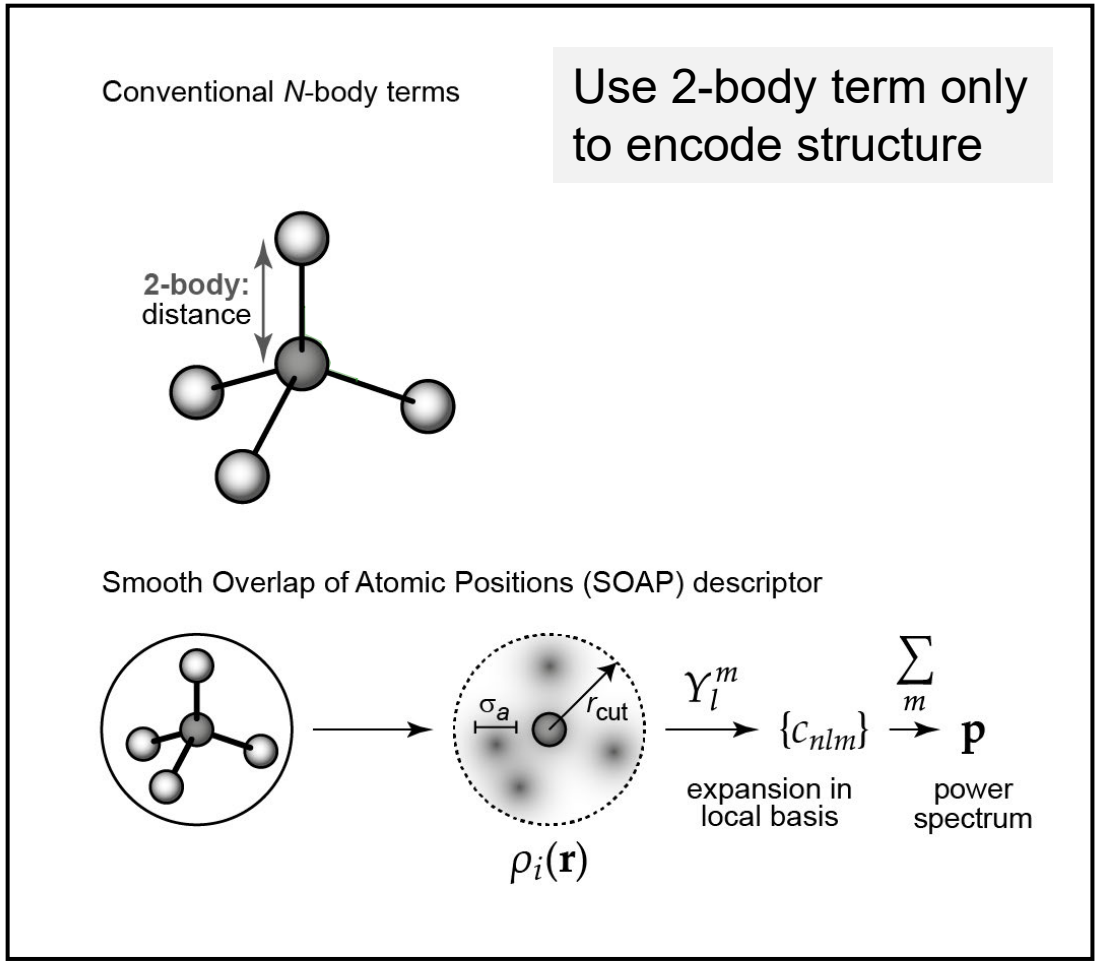
VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

SOAP: A. P. Bartók et al., *Phys. Rev. B* **2013**, *87*, 184115

Figure from: VLD et al., *Chem. Rev.* **2021**, *121*, 10073

Representations (*descriptors*) in practice



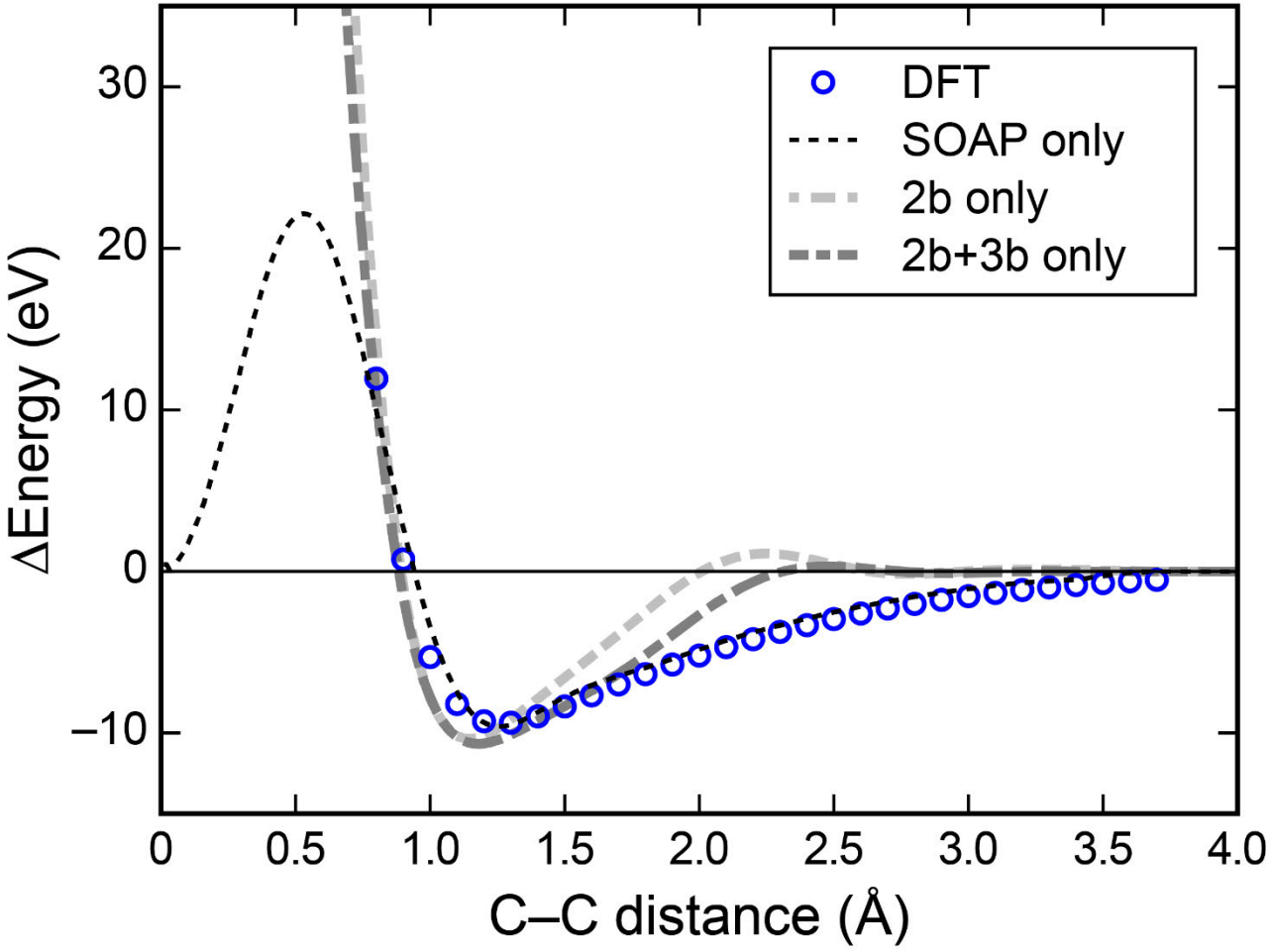


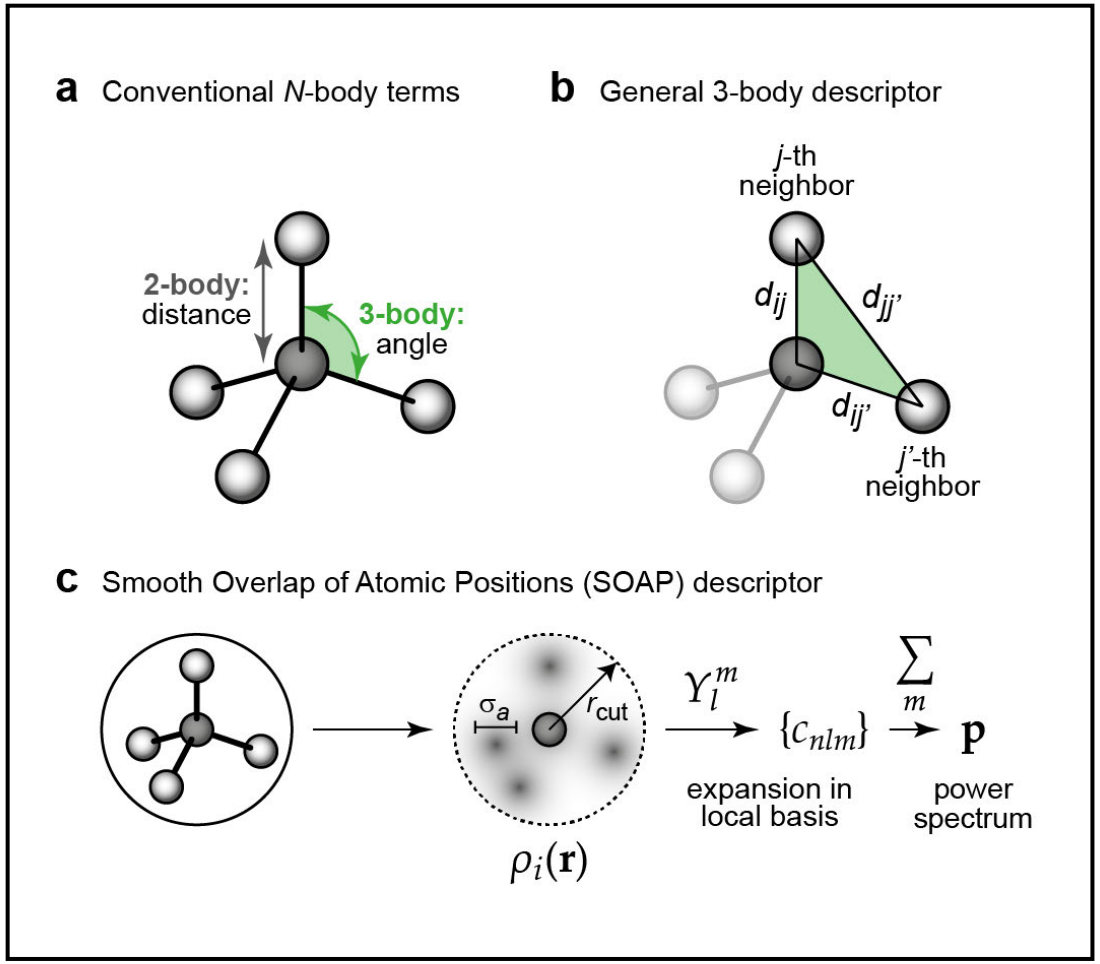
VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

SOAP: A. P. Bartók et al., *Phys. Rev. B* **2013**, *87*, 184115

Figure from: VLD et al., *Chem. Rev.* **2021**, *121*, 10073

Representations (*descriptors*) in practice



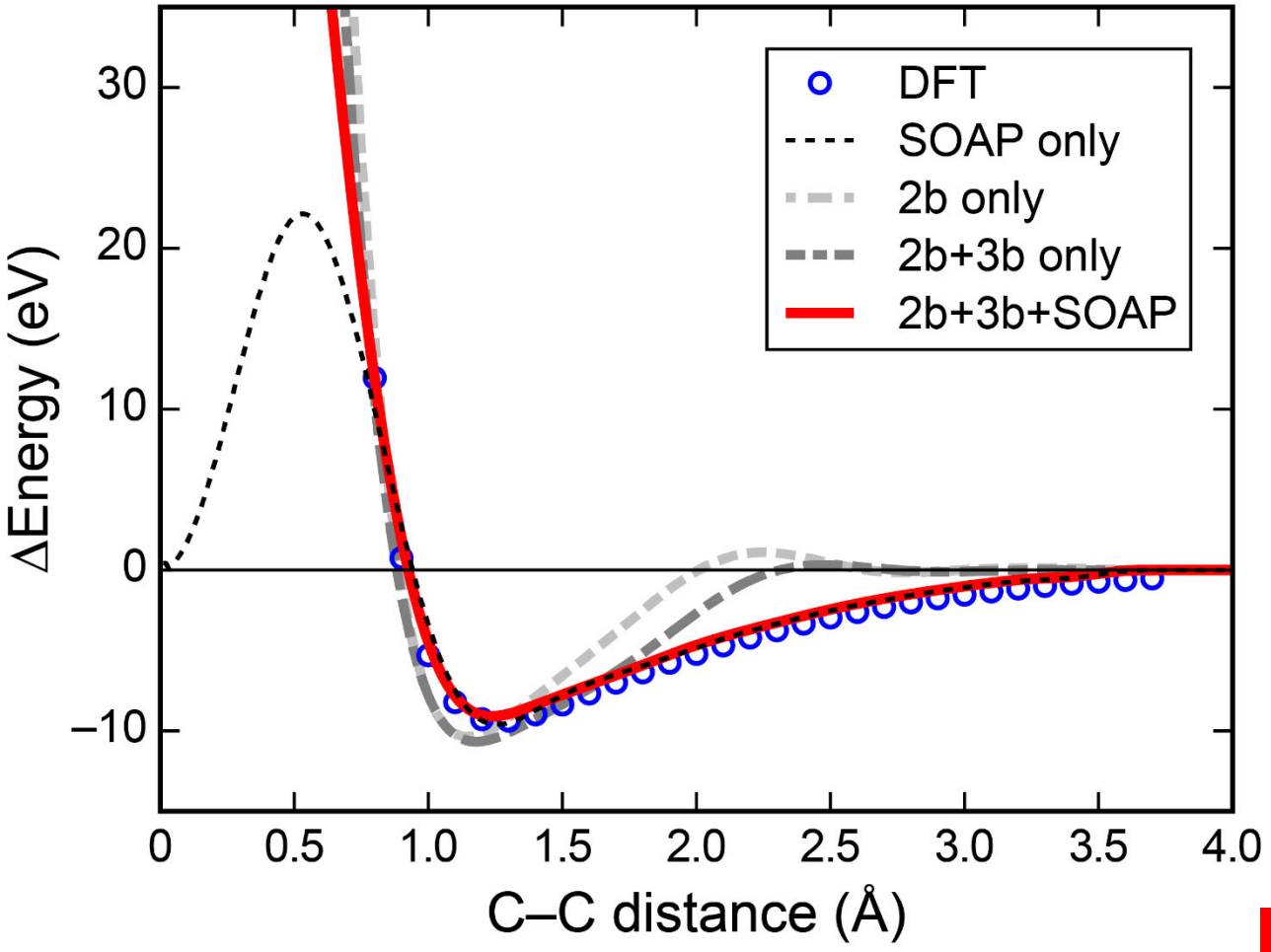


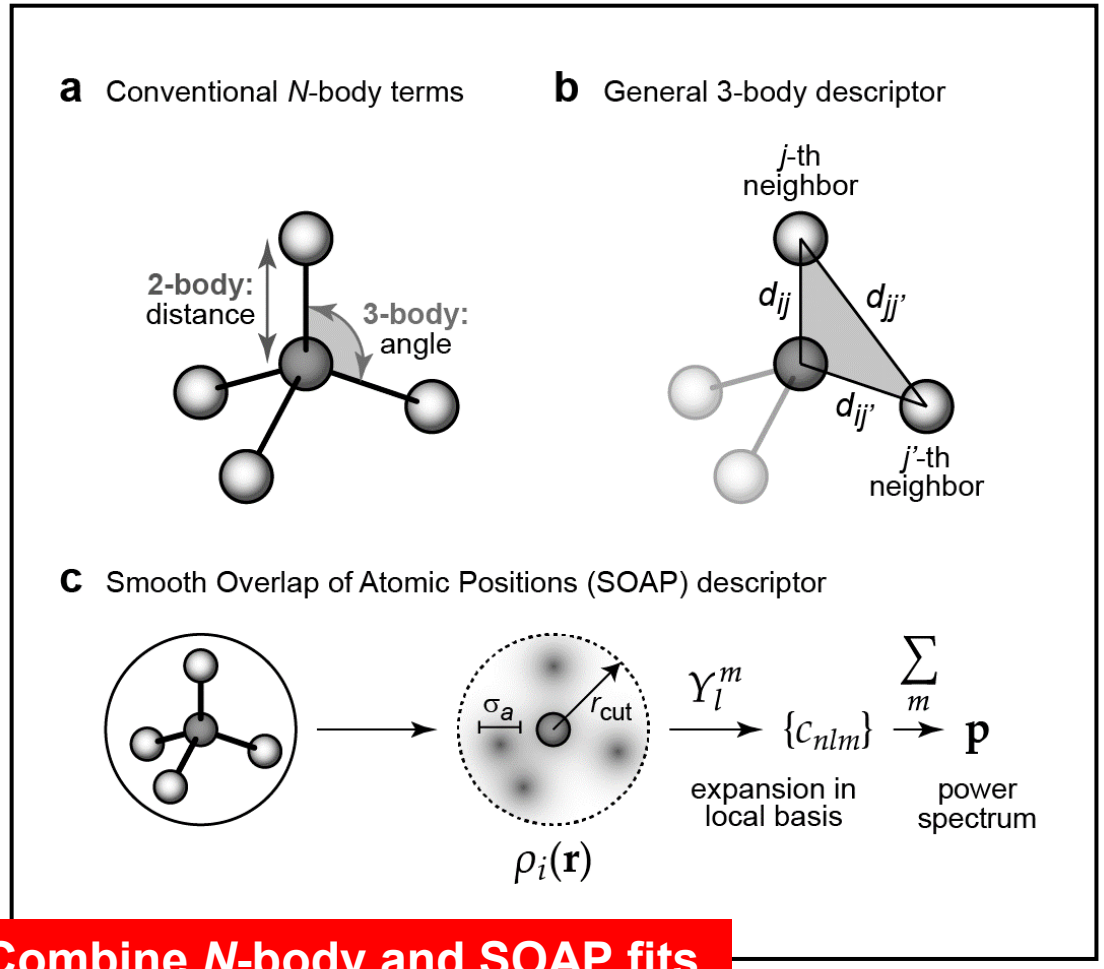
VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

SOAP: A. P. Bartók et al., *Phys. Rev. B* **2013**, *87*, 184115

Figure from: VLD et al., *Chem. Rev.* **2021**, *121*, 10073

Representations (*descriptors*) in practice



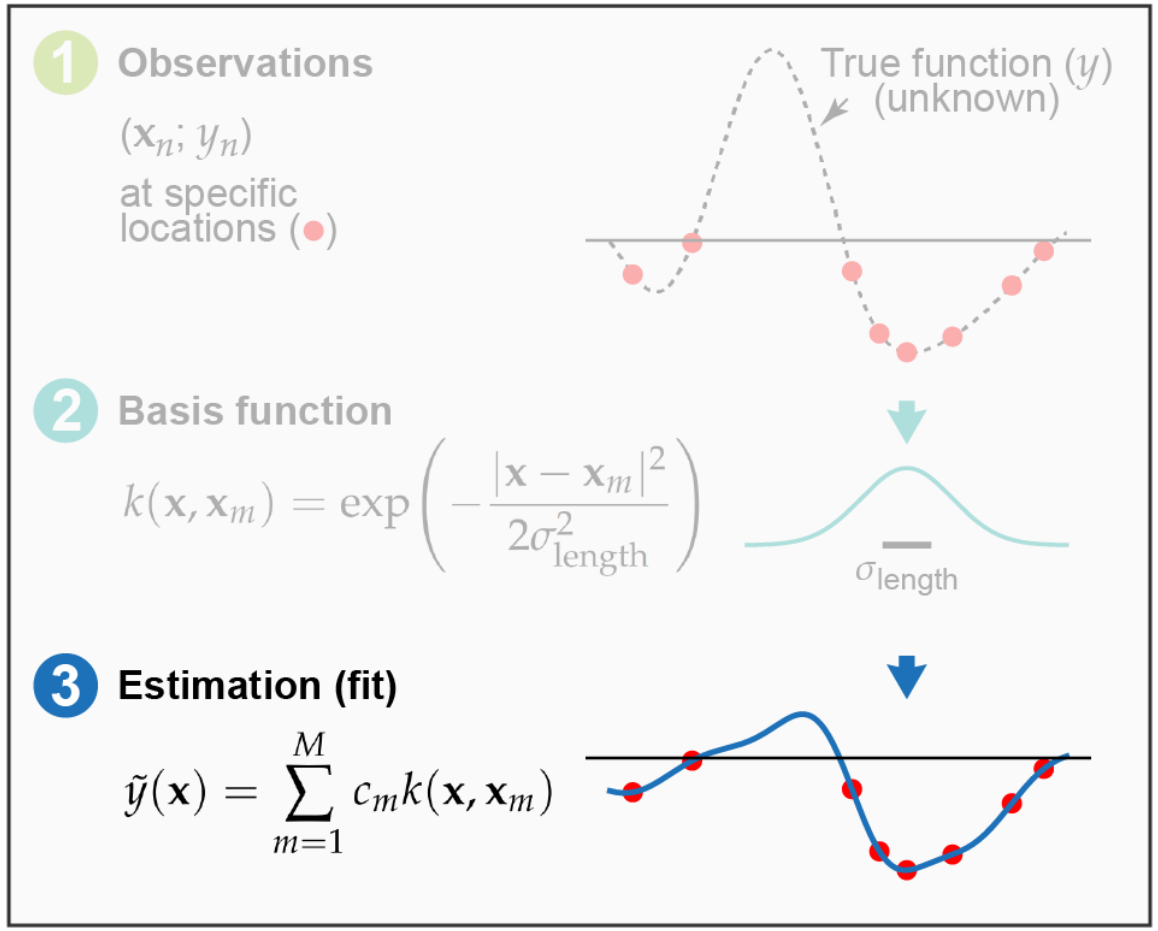


Combine *N*-body and SOAP fits

for **robustness** where needed, & **accuracy** where possible

VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

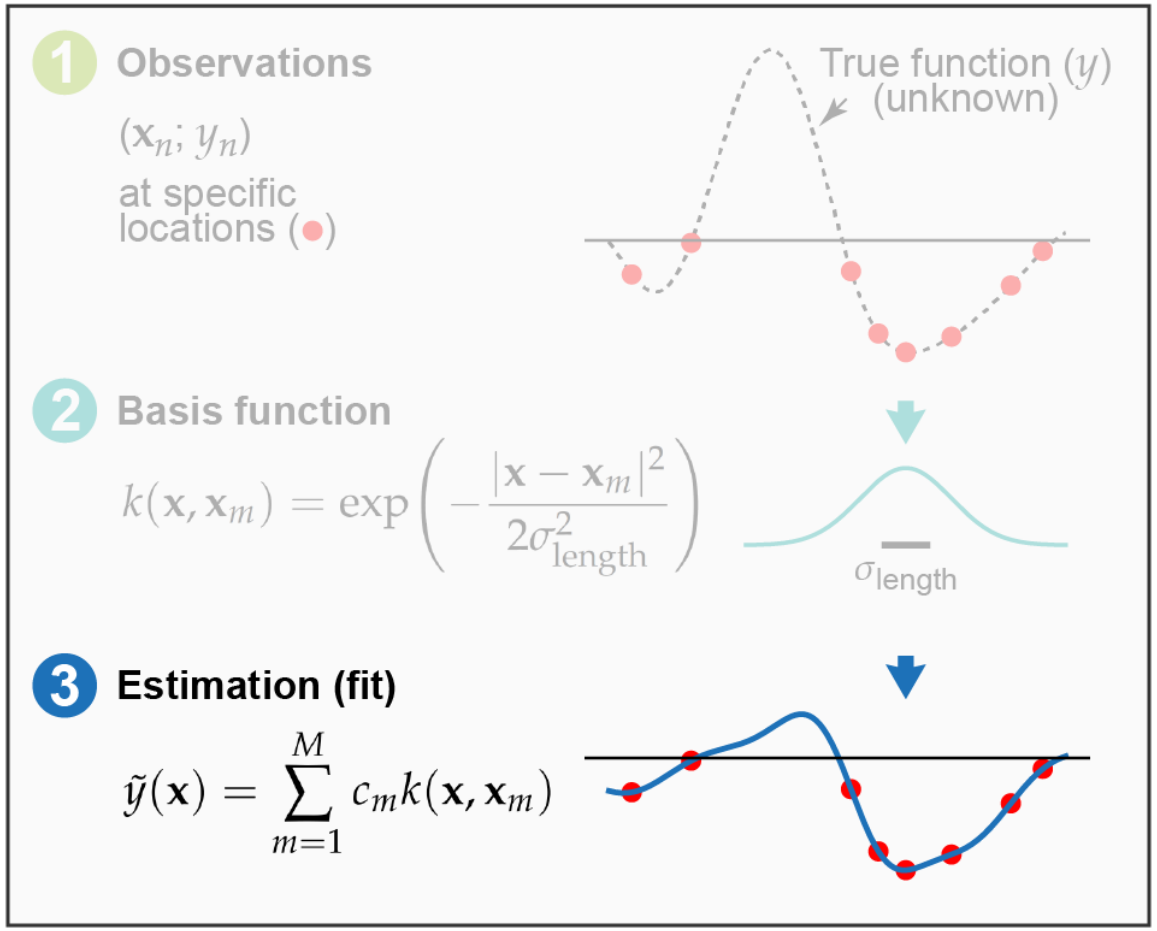
Gaussian process regression (GPR)

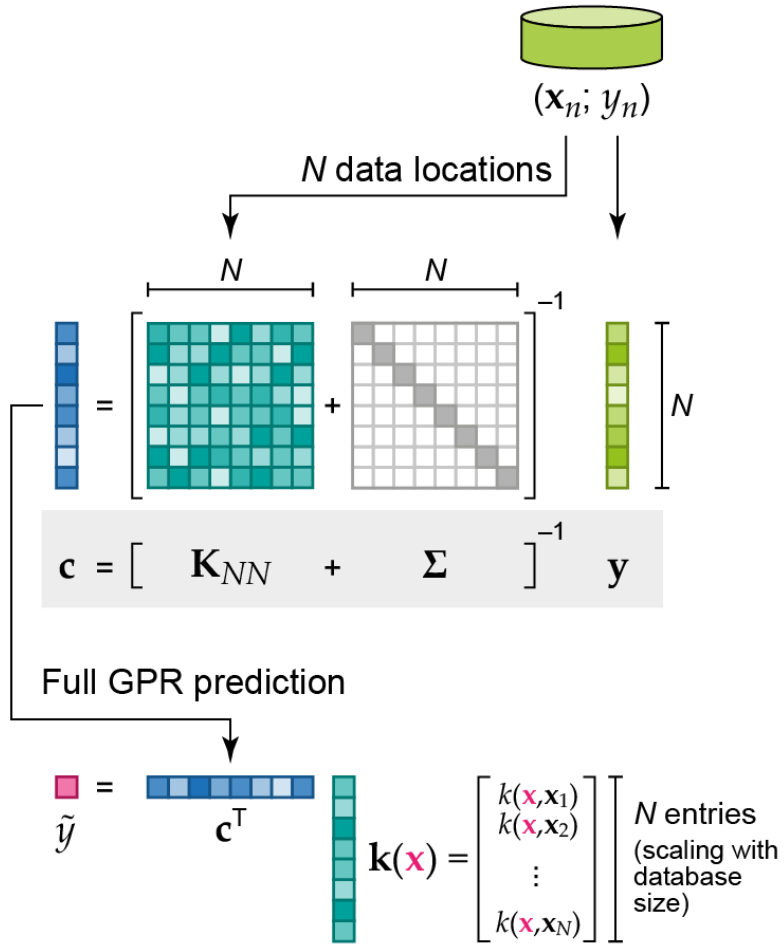


- **Observations**, using the "ground truth", make up the training data
- In atomistic ML: **x** encodes atomic environment, *y* can be any per-atom property
- **Basis functions** are built using a kernel similarity measure, *k*
- In atomistic ML: from simple pair distance to more complex forms
- **Estimation** means predicting *y* at a new, unknown location **x**
- To do this, we need *k* and pre-determined fitting coefficients, *c*

Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR)



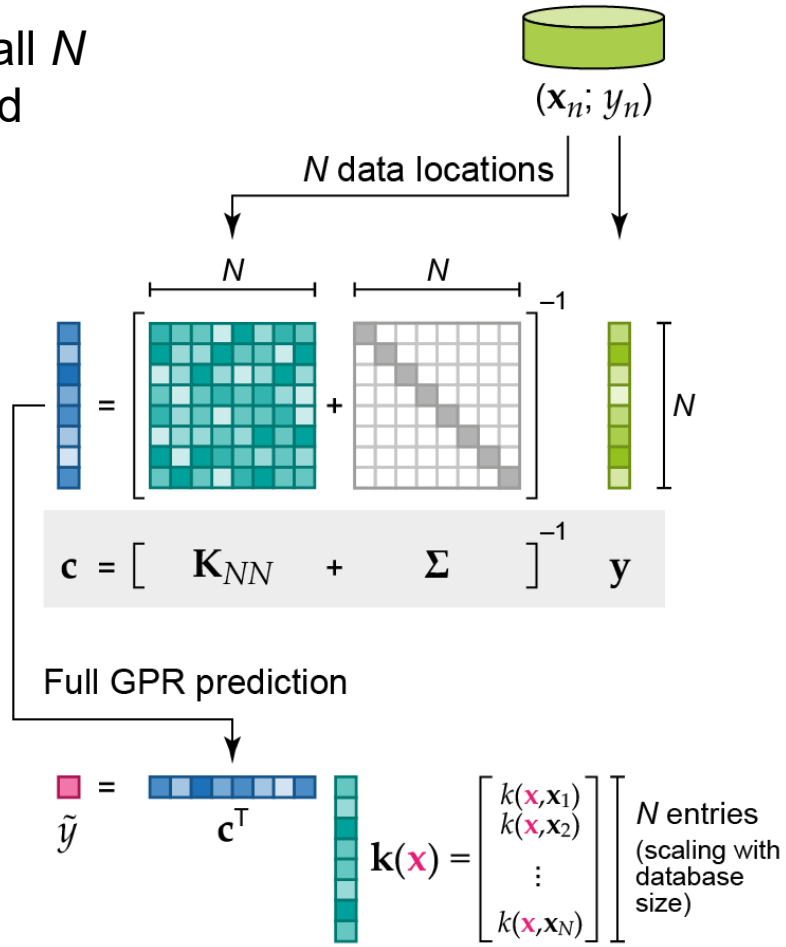


Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR)

- Finding the coefficients, *c*, means solving a linear algebra problem
- Σ is a **regulariser** that encodes the expected noise in the input data
- (effect of regularisation: see examples at https://arxiv.org/abs/2211.16443)

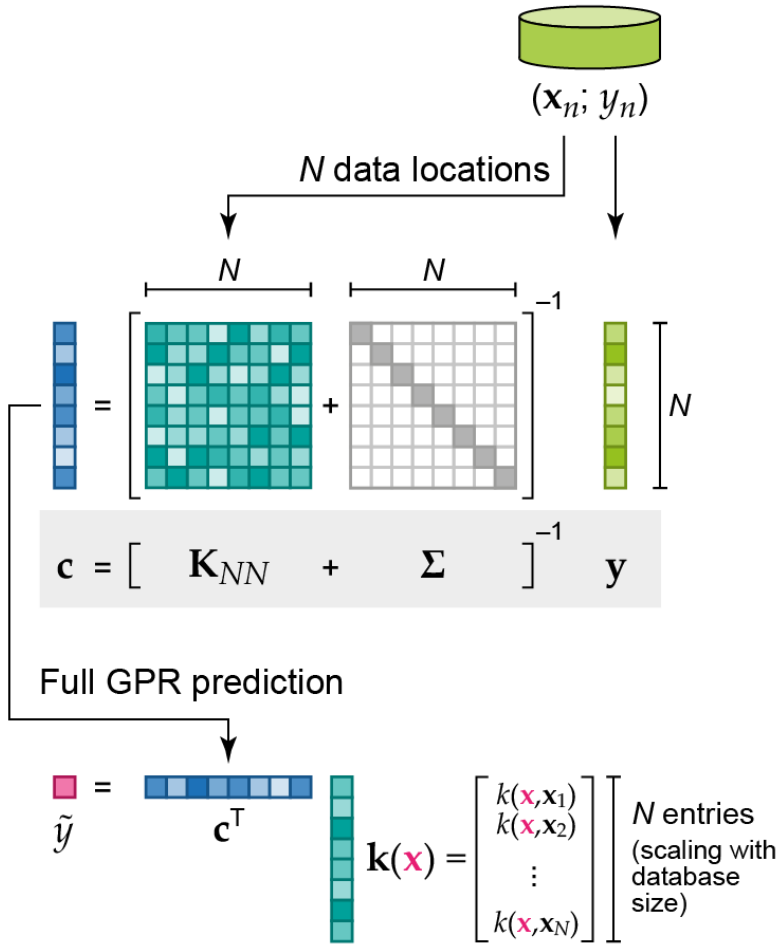
- This is "**full GPR**": all *N* data points are used



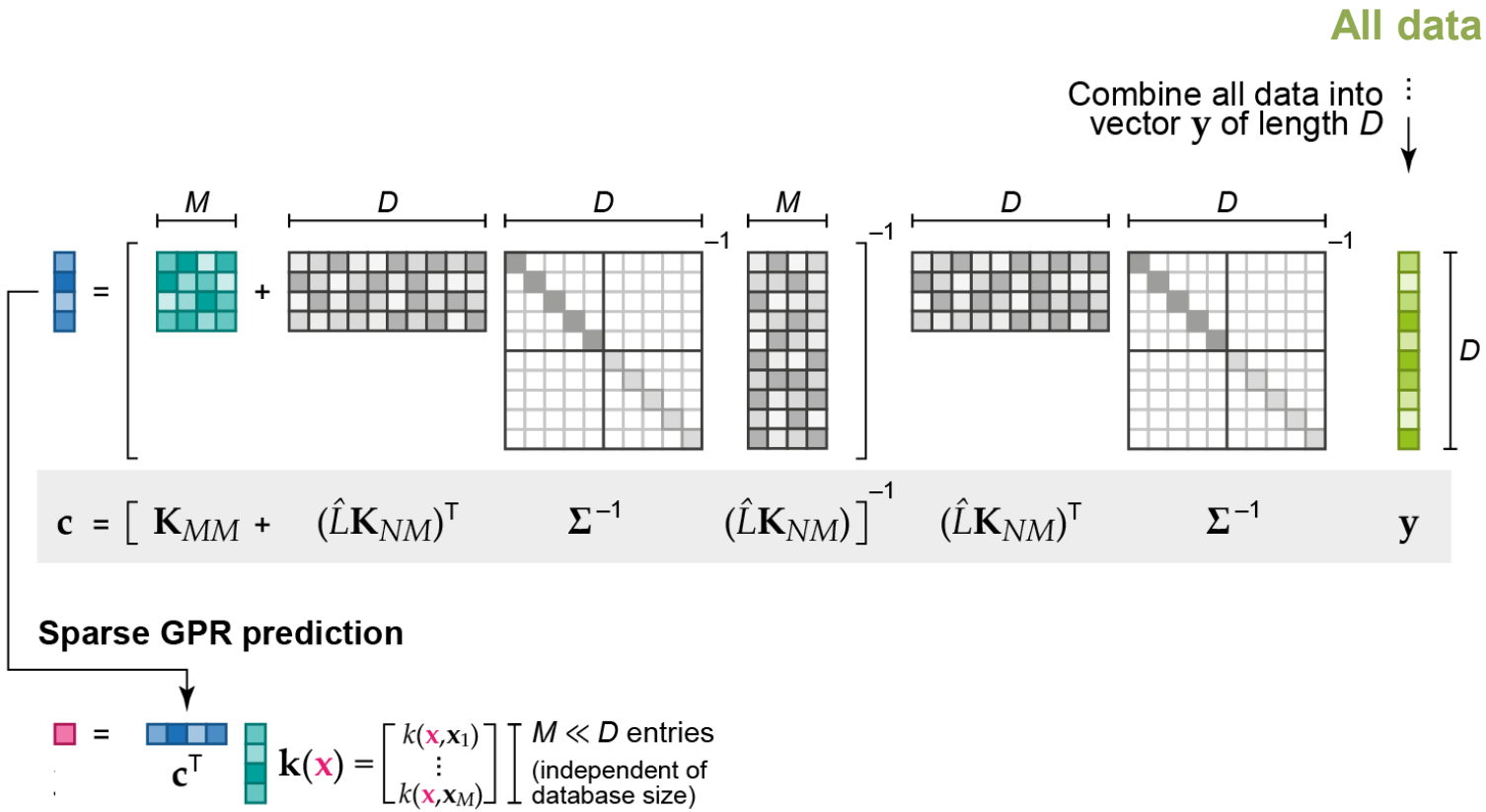
Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR)

a Full GPR fitting

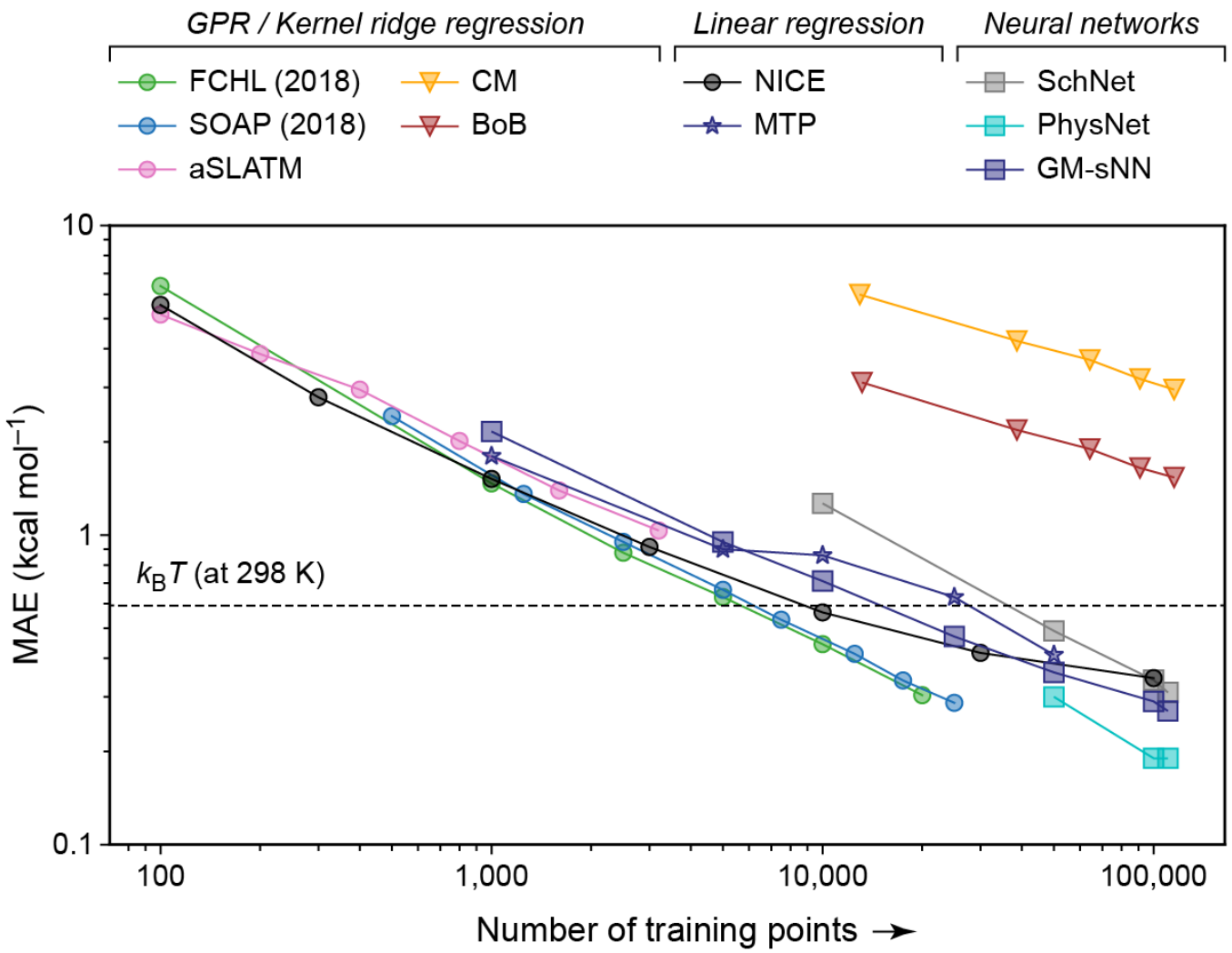


b Sparse GPR fitting: e.g., in **GAP models**



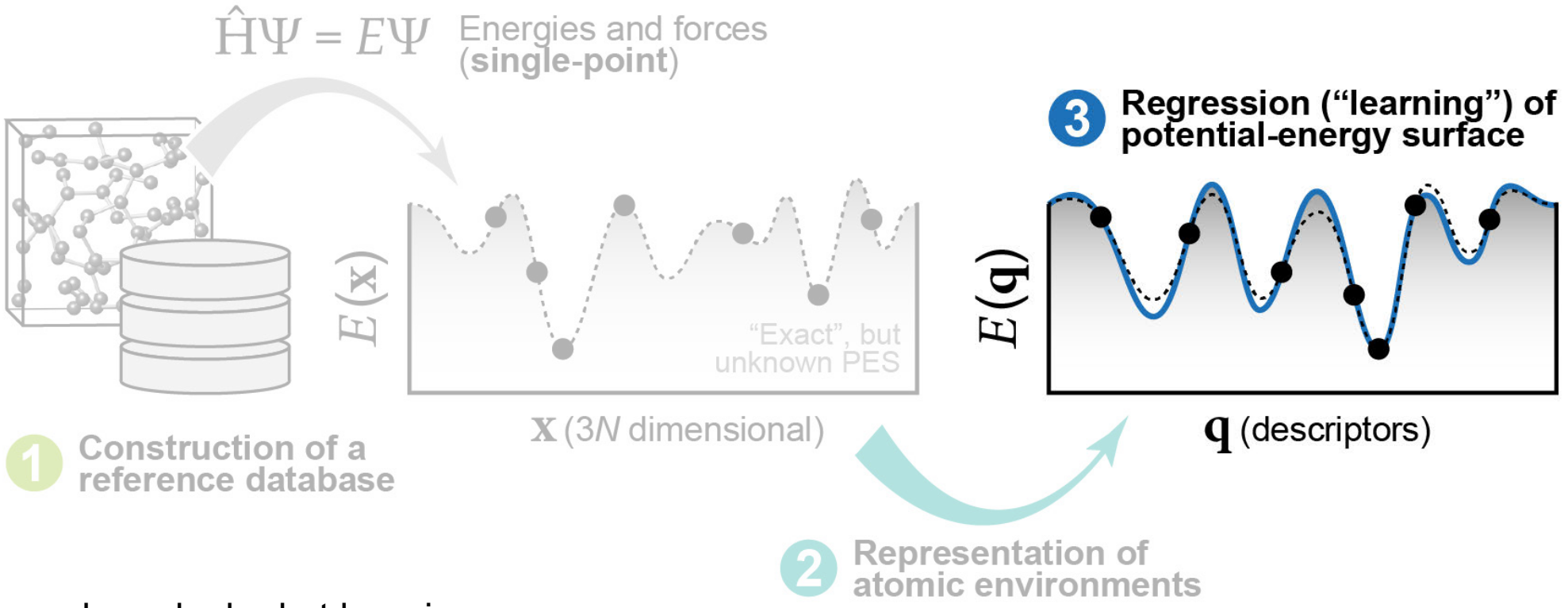
Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR)



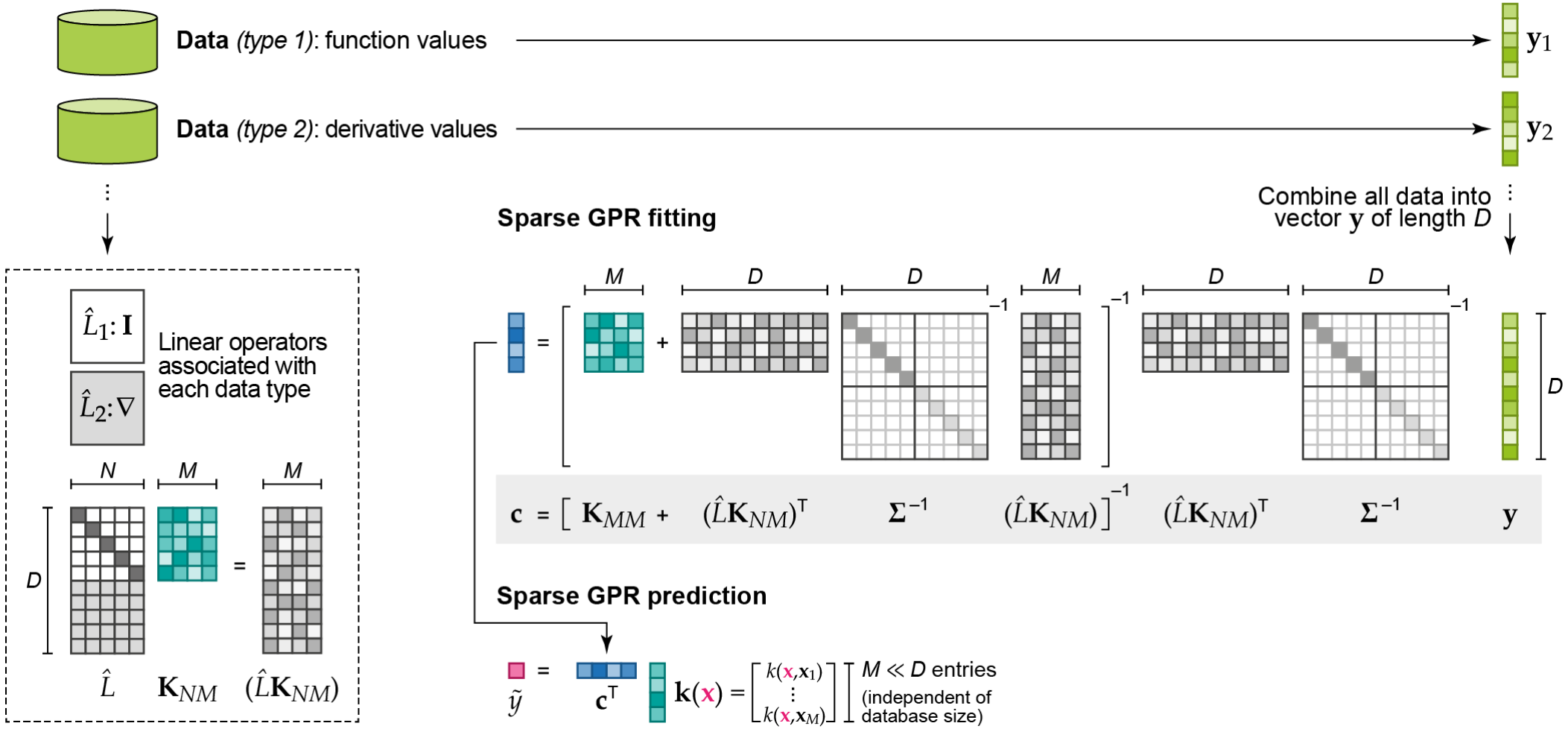
Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian process regression (GPR) for interatomic potentials



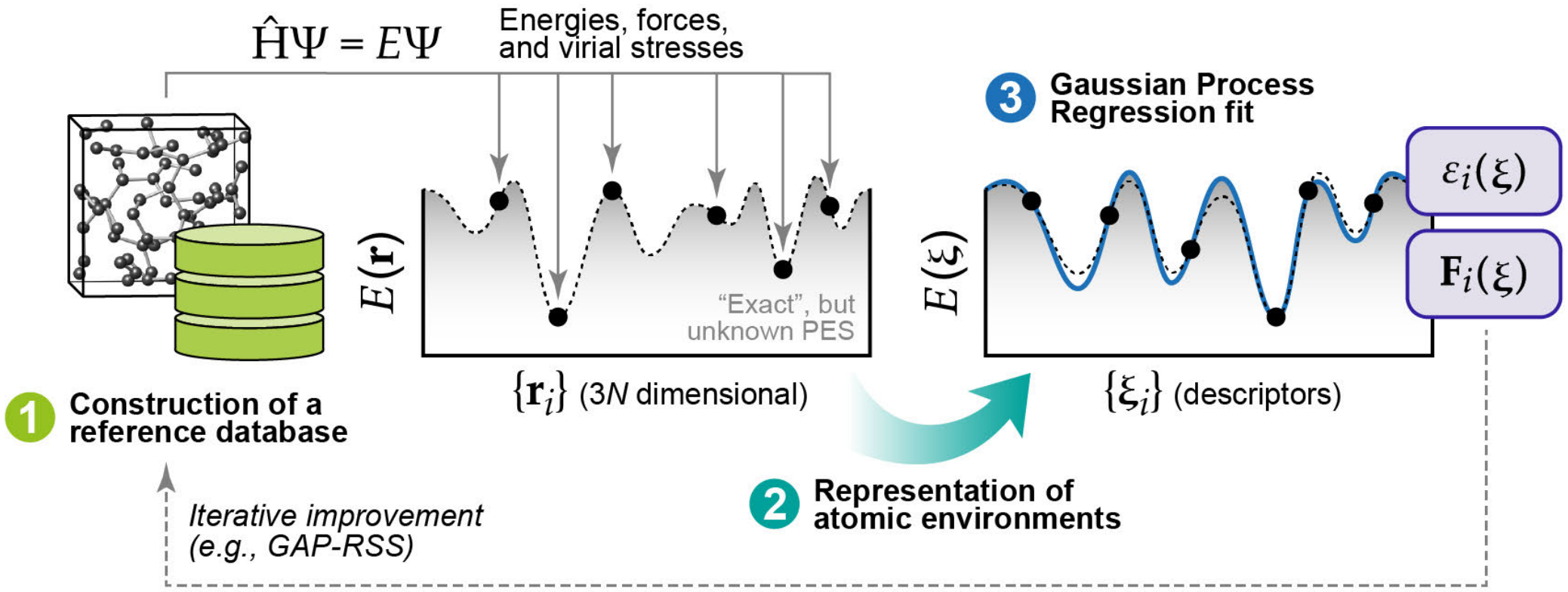
- So far, we have looked at learning an atomistic property, *y*, itself
- For interatomic potentials: much of the training data are forces (**derivatives**)

Gaussian process regression (GPR) for interatomic potentials



Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian approximation potential (GAP) models



Chem. Rev. **2021**, *121*, 10073: "Gaussian Process Regression for Materials and Molecules" **(Open Access)**

Gaussian approximation potential (GAP) models

Reference data

- All available data are used: total energy, forces, and stresses (for periodic systems), combined into a single ML fit. The design of the input database is critical to the success of the model and has been a cornerstone of all presently available general-purpose GAPs. The selection of reference data is as much an area of ongoing methods development as is that of representation and regression (section 4.1).

The fit itself

- Hyperparameters of the GAP model are chosen and fixed a priori as much as possible and optimized only where required. The main hyperparameters are (i) the relevant length scales, which define the cutoff radius and the smoothness of the kernel, and (ii) the expected errors (arising both from noise in the input data and limitations of the model, e.g., due to the necessarily finite cutoff radius; section 4.4), which determine the regularization of the fit (section 4.6).

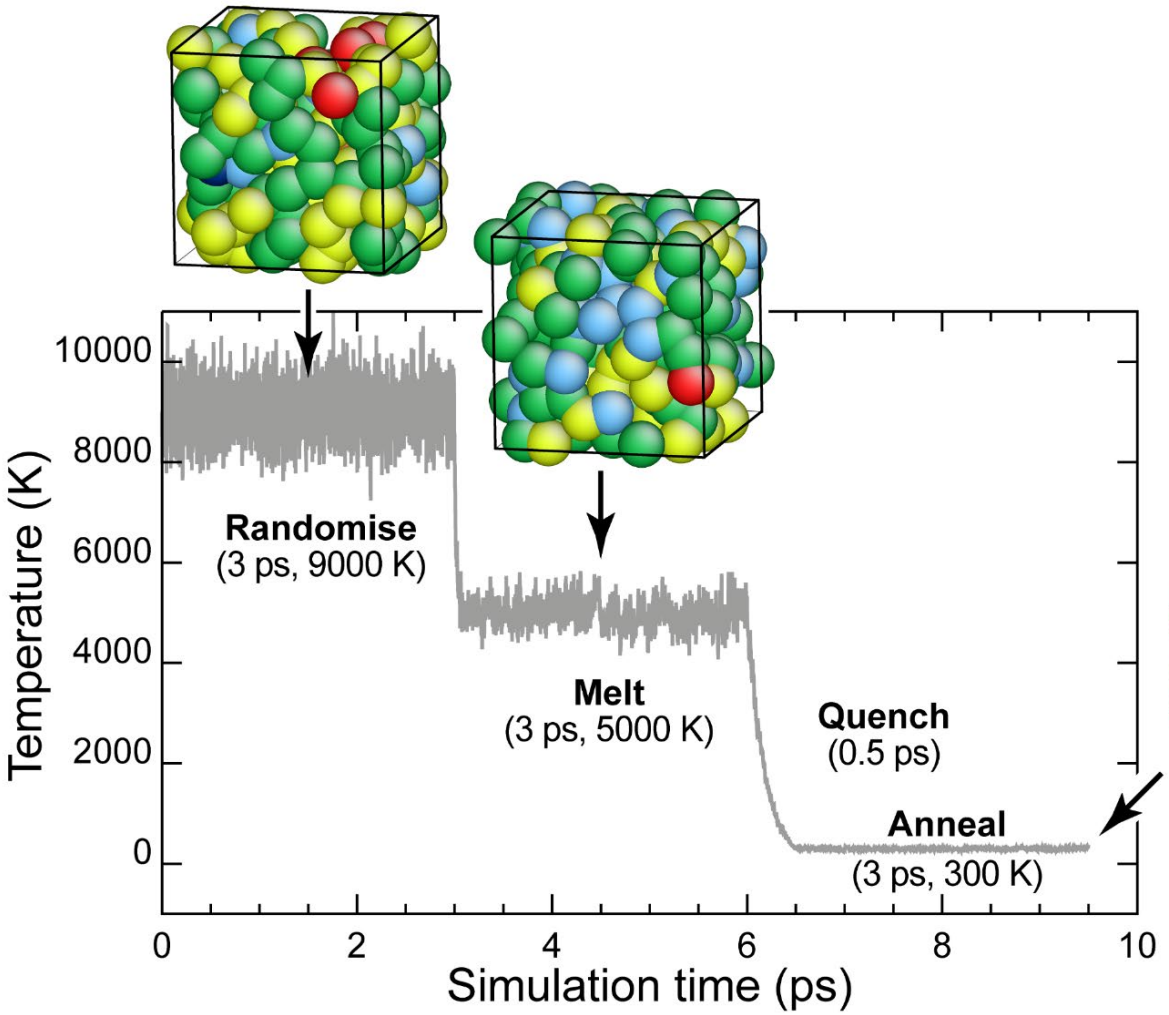
Representation

- The choice and specification of structural descriptors (representation) is tightly coupled with the choice of kernels, and both are an essential part of the user input. They incorporate prior knowledge about the nature of the potential-energy function—specifically, its regularity. Commonly used examples are distances and angles between atoms together with a squared exponential (Gaussian) kernel, or the many-body SOAP representation with a polynomial kernel. These are not mutually exclusive: low-dimensional kernel models can be fitted together with many-body ones, with appropriate weighting between them. All representations and kernels in GAP have finite distance cut-offs, typically about 5−6 Å, and therefore they represent the local environments of the atoms (section 4.2).

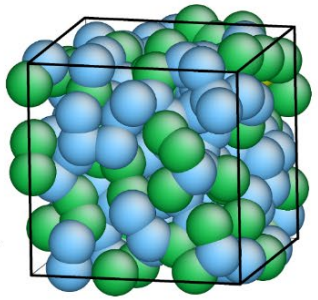
Available for non-commercial research at **https://github.com/libAtoms/QUIP**

Chem. Rev. **2021**, *121*, 10073: “Gaussian Process Regression for Materials and Molecules” **(Open Access)**

Gaussian approximation potential (GAP) models

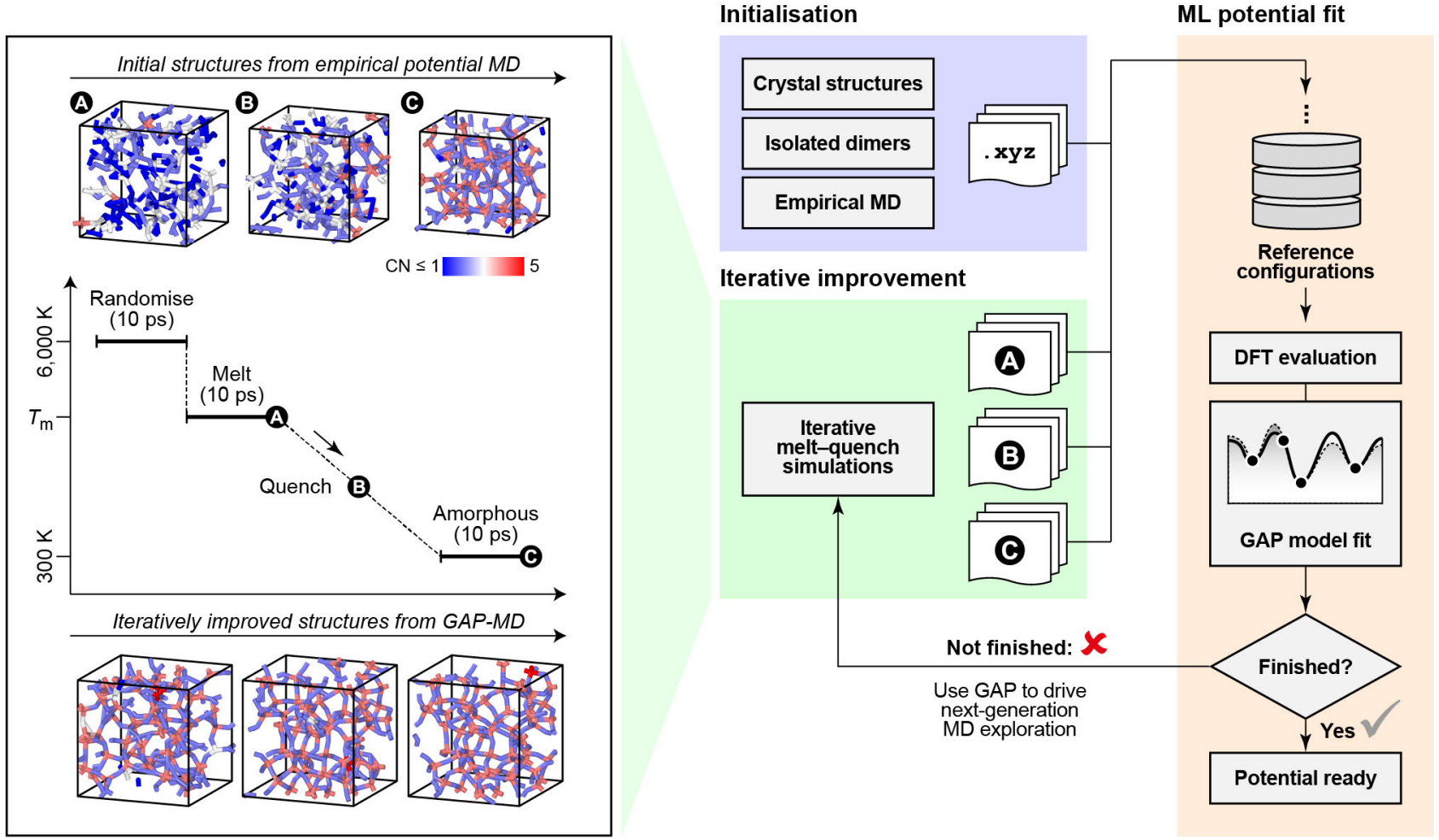


Tutorial: **melt–quench simulations** using GAP interfaced to LAMMPS



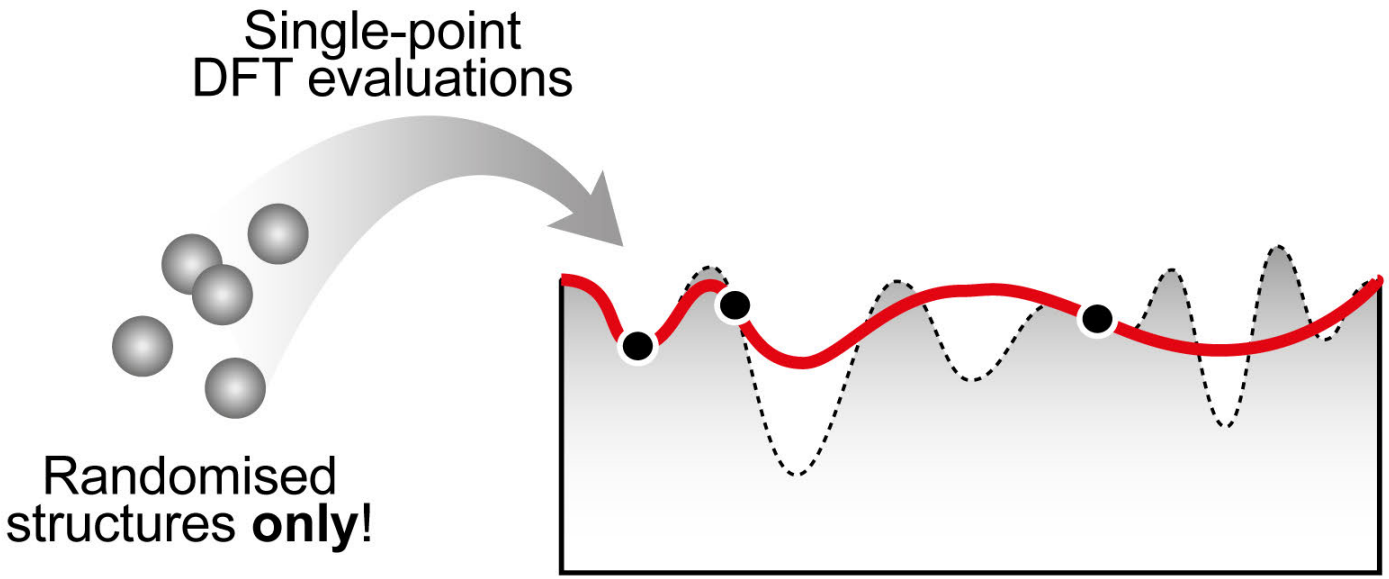
VLD, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203

Reference data are critical

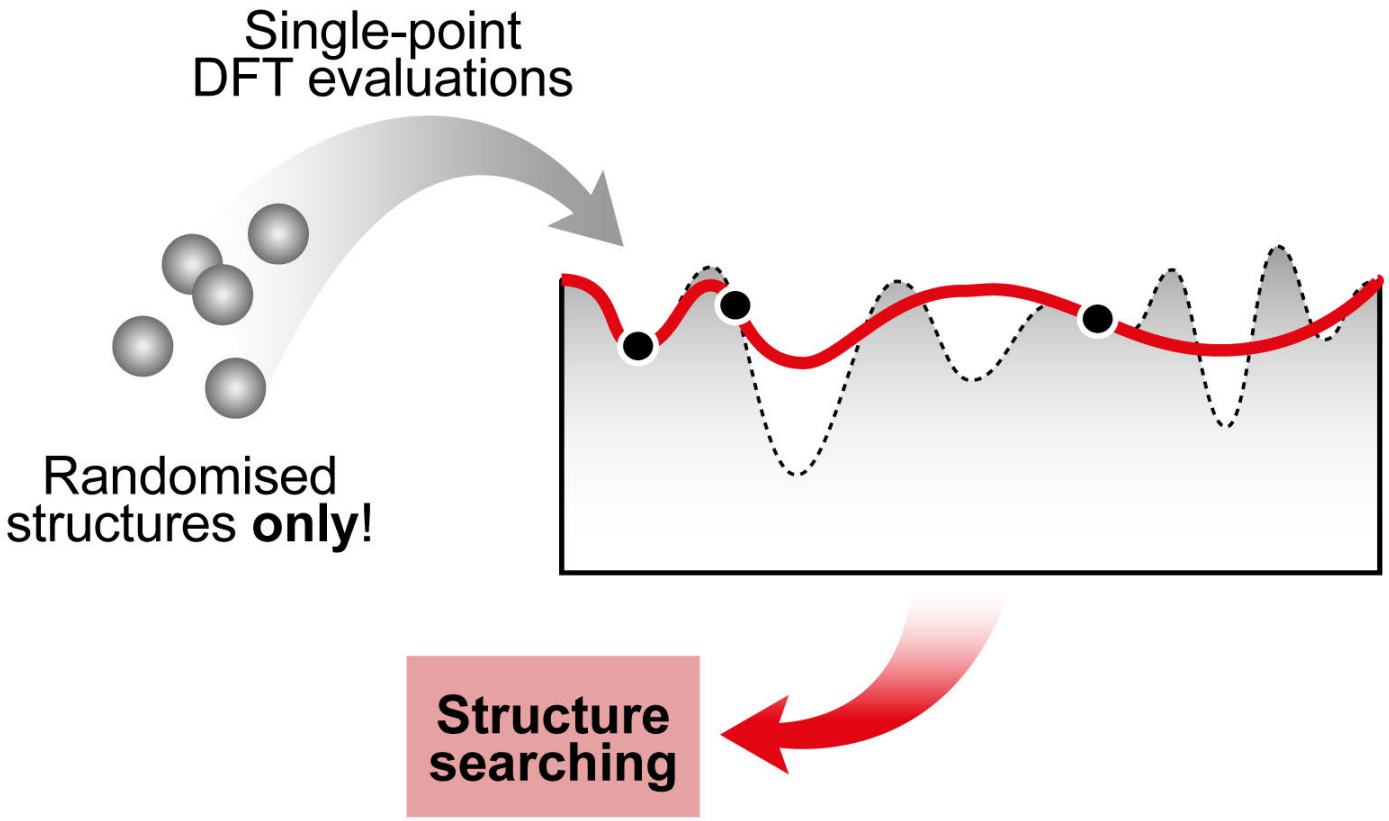


L. C. Erhard, J. Rohrer, K. Albe, VLD, *npj Comput. Mater.* **2022**, *8*, 90

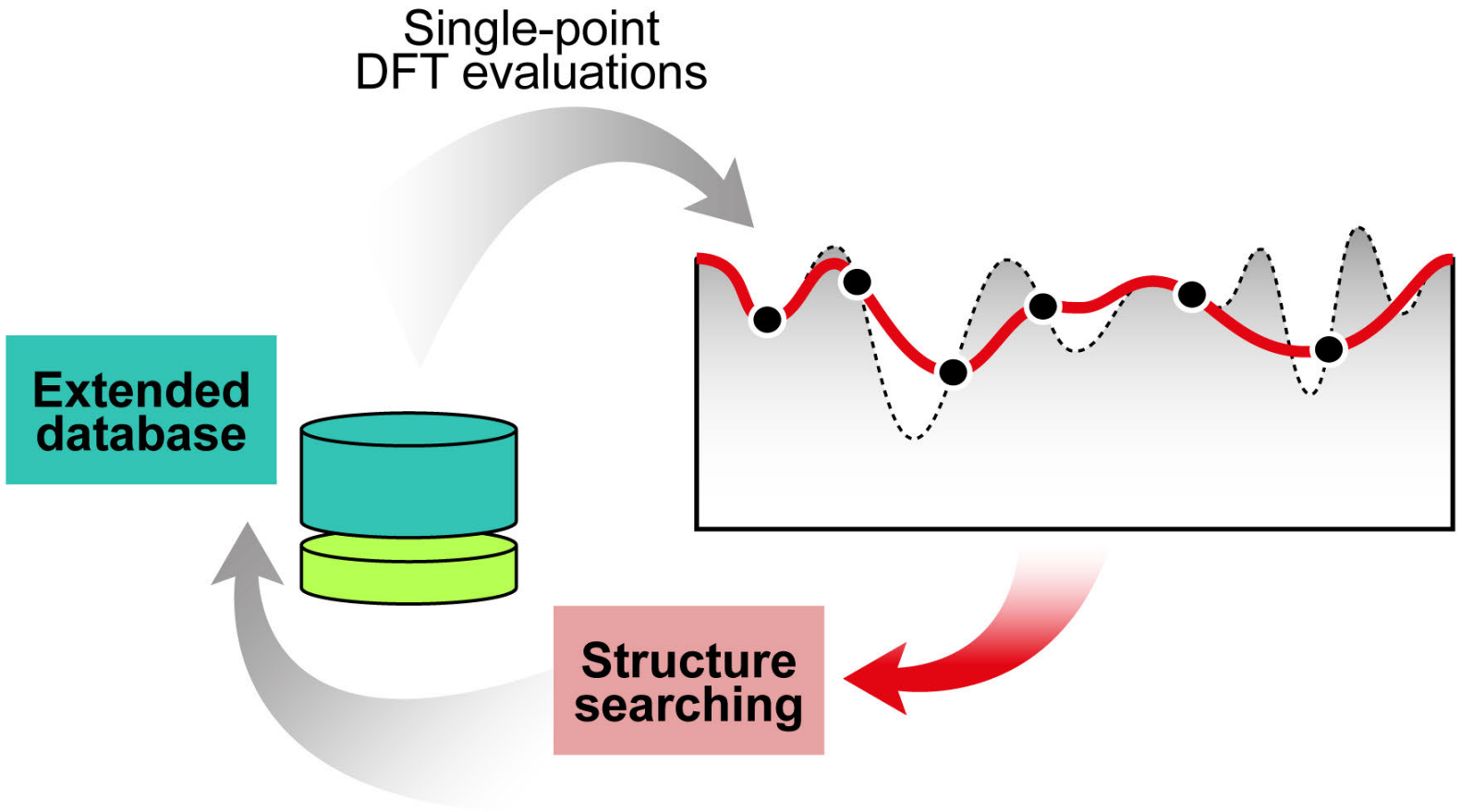
De novo exploration & fitting



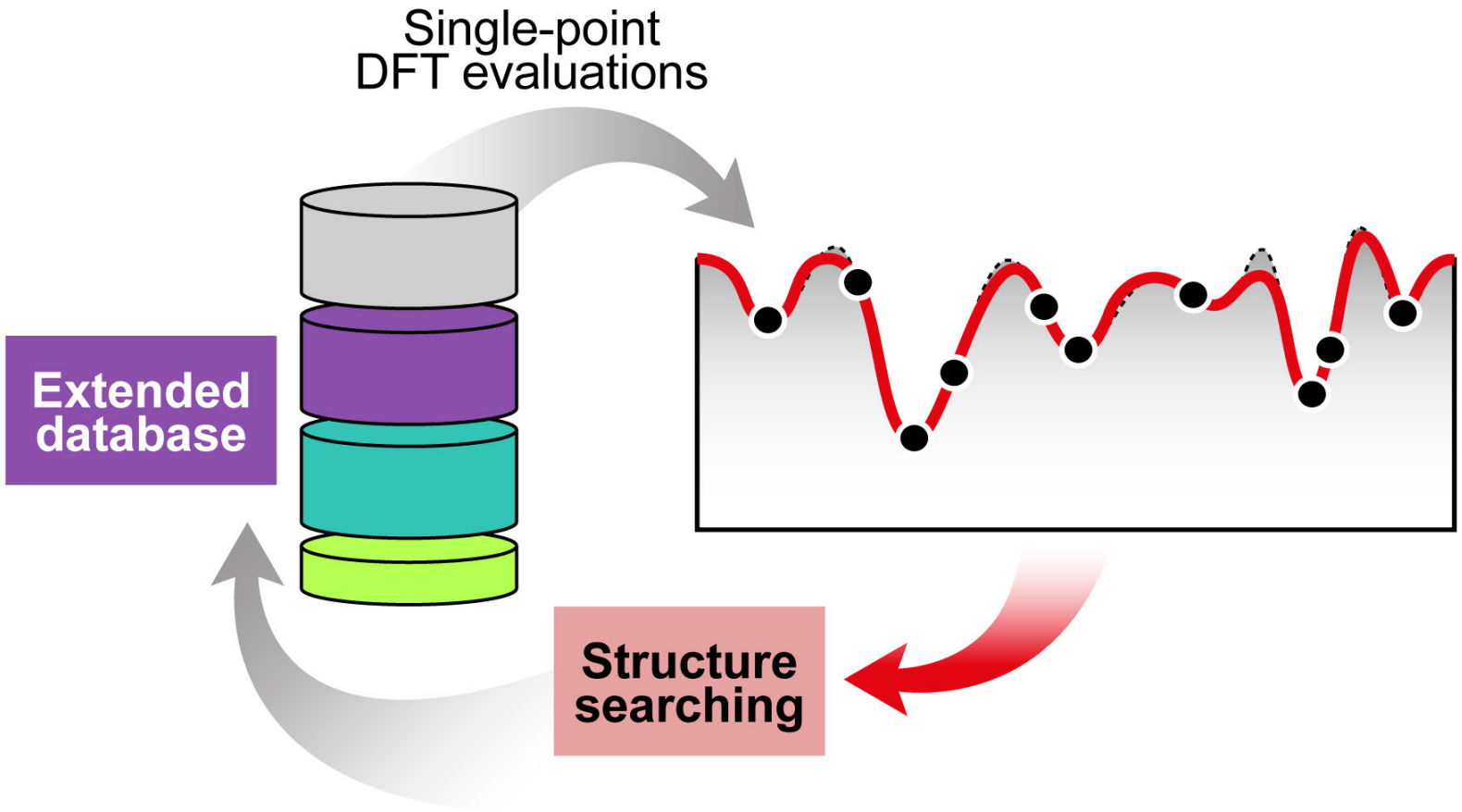
De novo exploration & fitting



De novo exploration & fitting

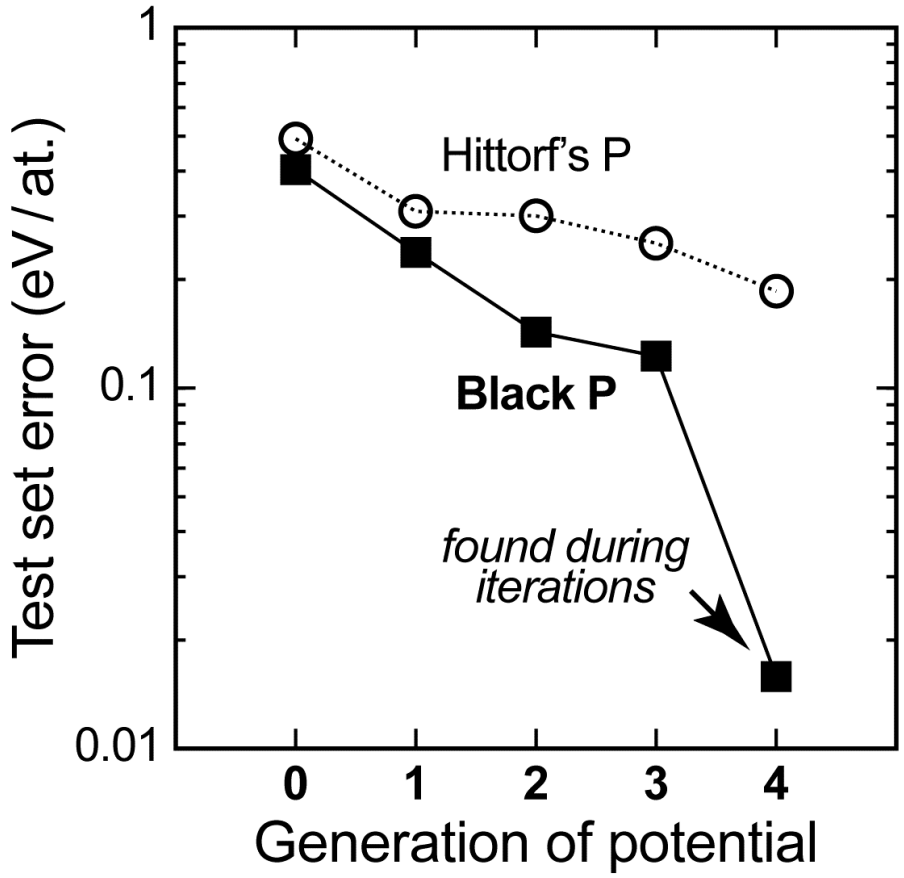


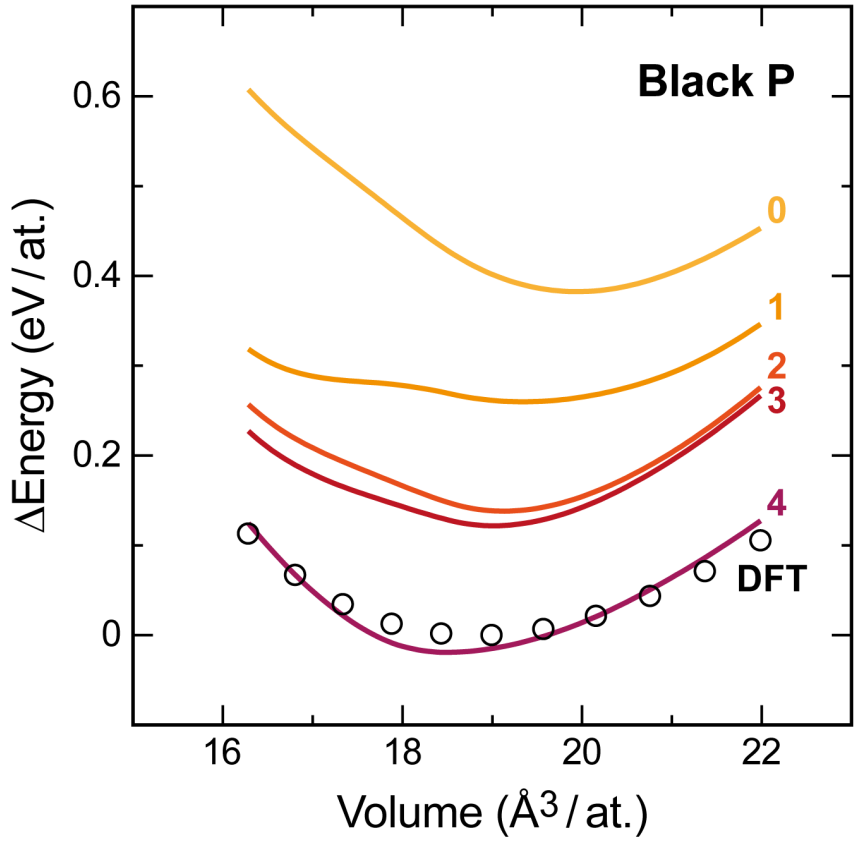
De novo exploration & fitting



VLD, C. J. Pickard, G. Csányi, *Phys. Rev. Lett.* **2018**, *120*, 156001
VLD, D. M. Proserpio, G. Csányi, C. J. Pickard, *Faraday Discuss.* **2018**, *211*, 45

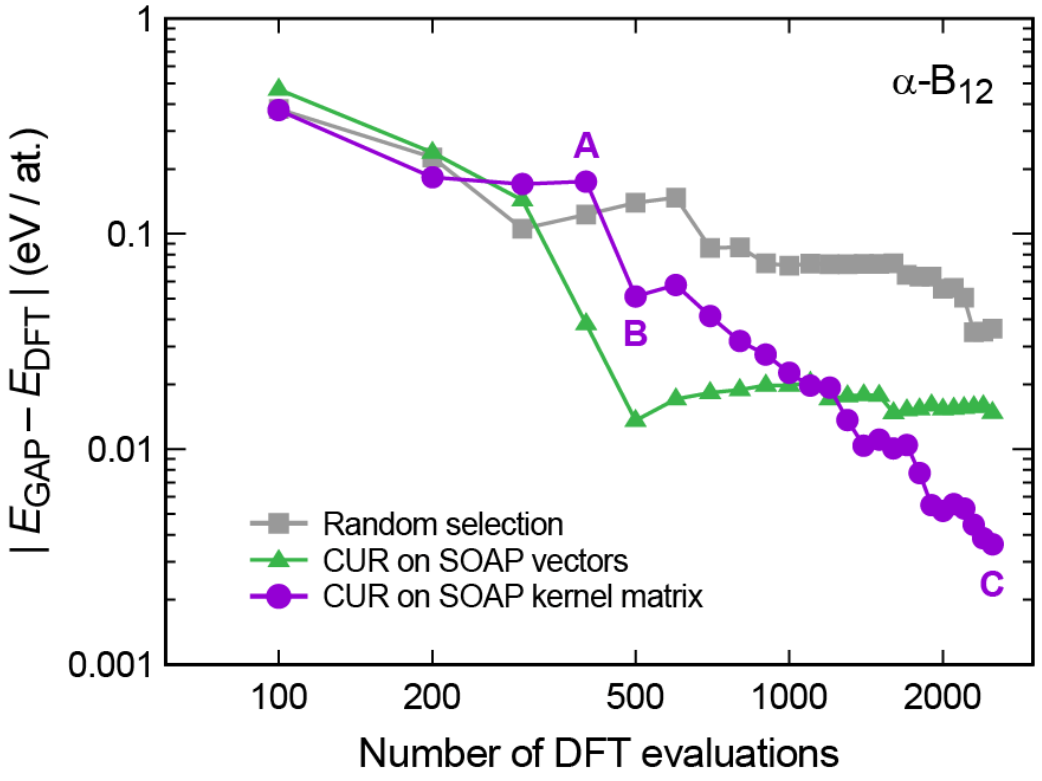
De novo exploration & fitting

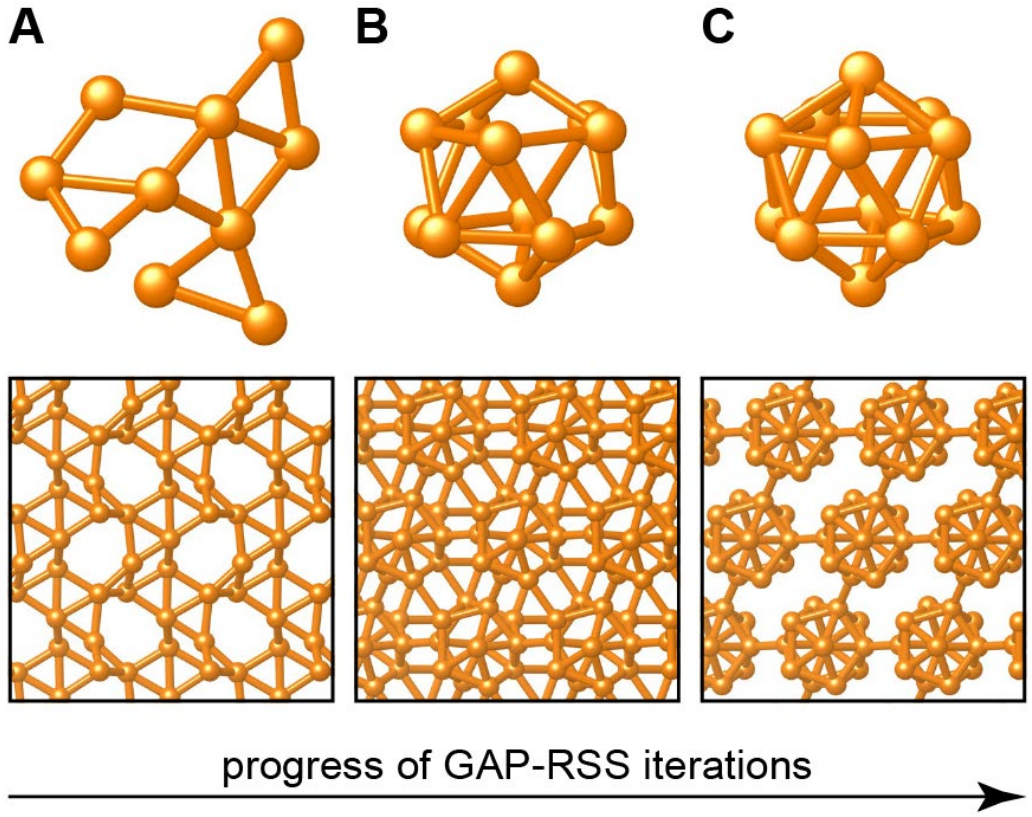




VLD, C. J. Pickard, G. Csányi, *Phys. Rev. Lett.* **2018**, *120*, 156001
VLD, D. M. Proserpio, G. Csányi, C. J. Pickard, *Faraday Discuss.* **2018**, *211*, 45

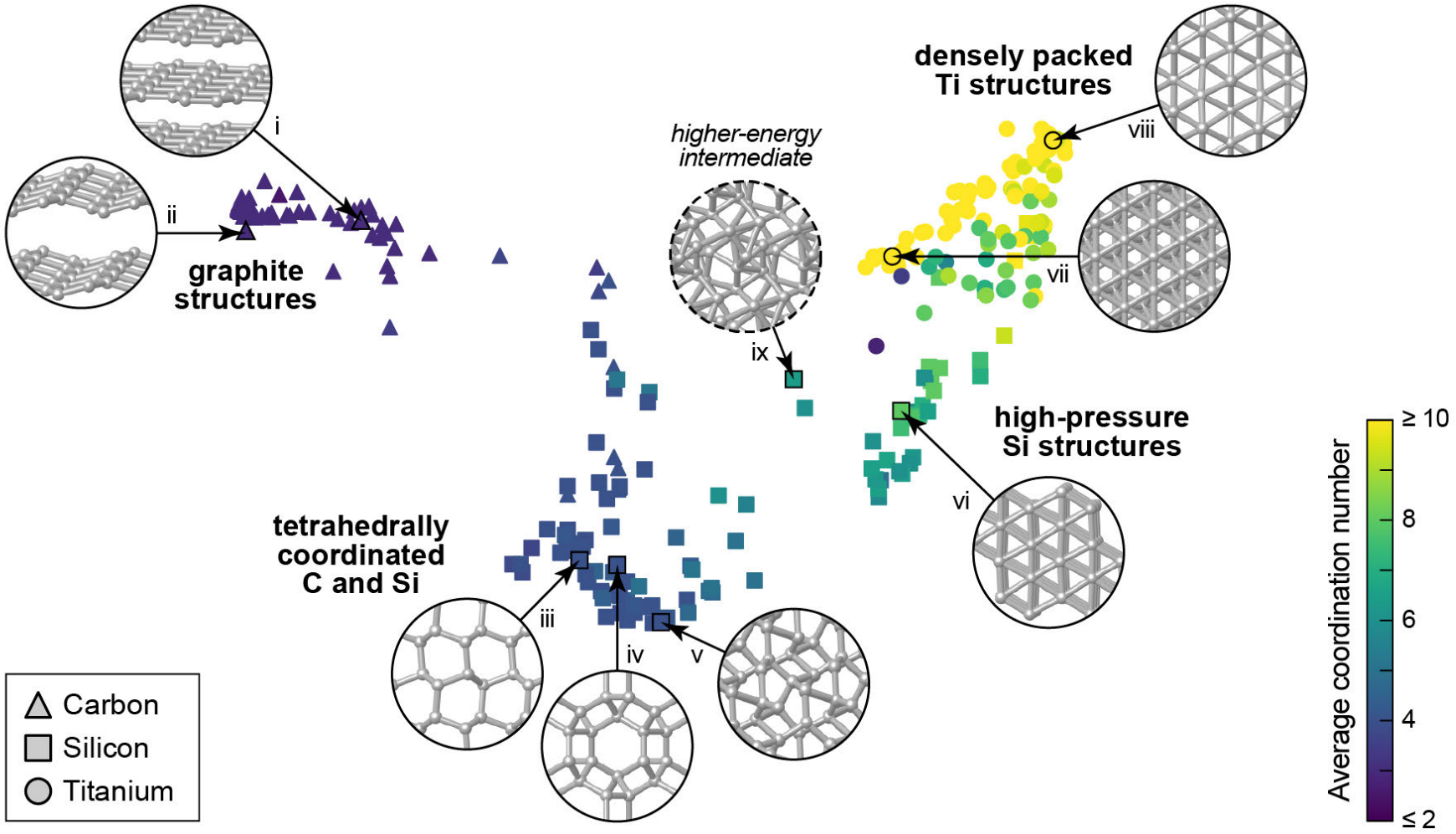
De novo exploration & fitting





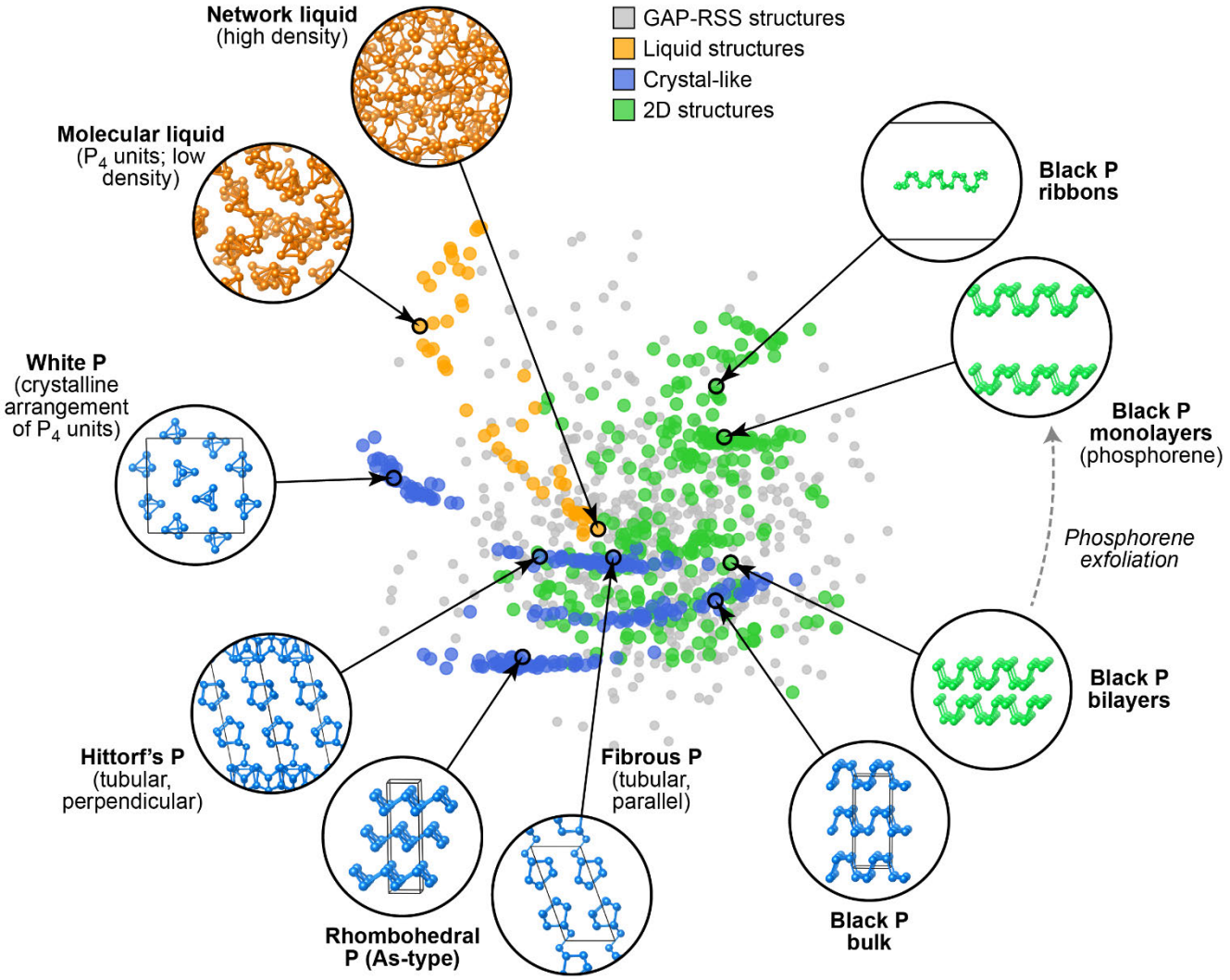
N. Bernstein, G. Csányi, VLD, *npj Comput. Mater.* **2019**, *5*, 99

De novo exploration & fitting



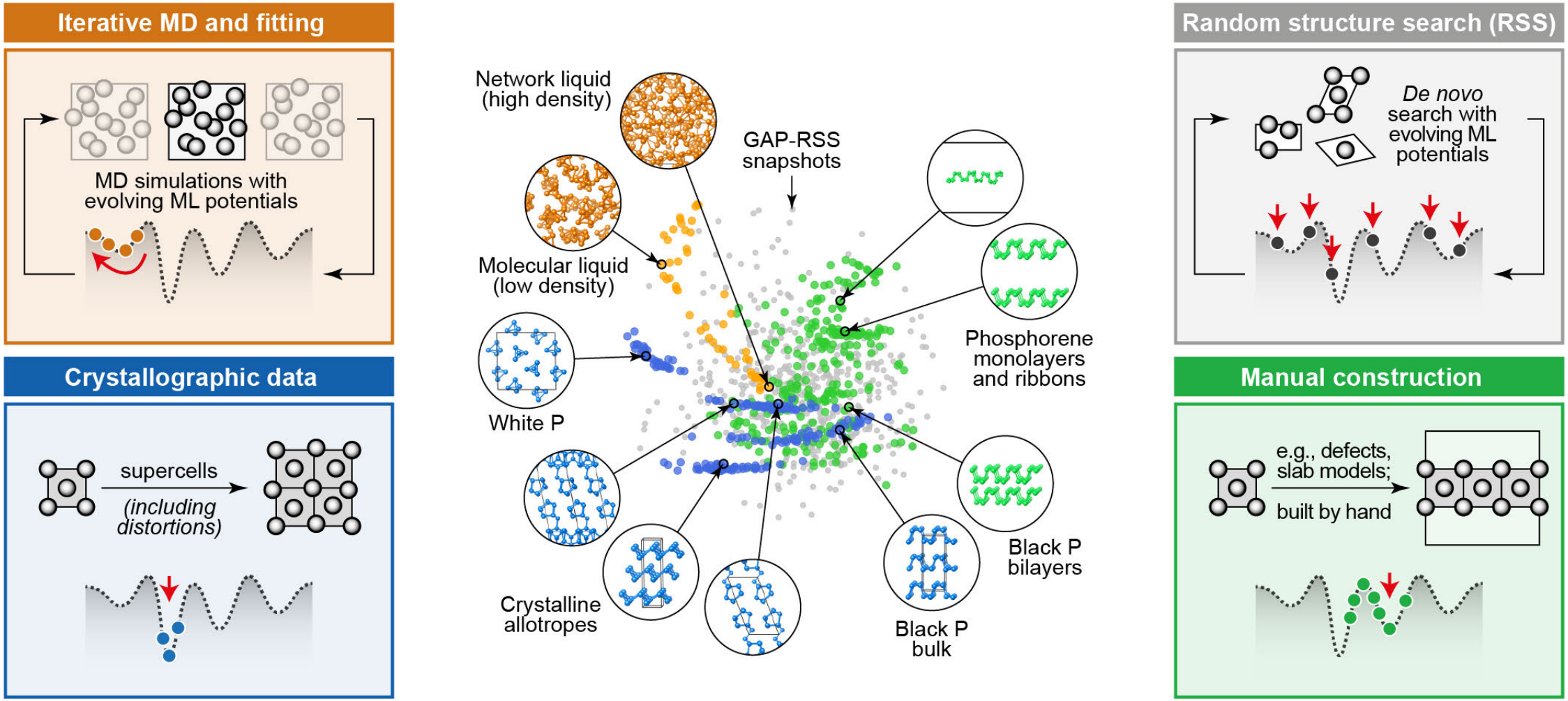
N. Bernstein, G. Csányi, VLD, *npj Comput. Mater.* **2019**, *5*, 99

GAP-RSS as *starting point* for ML potential databases



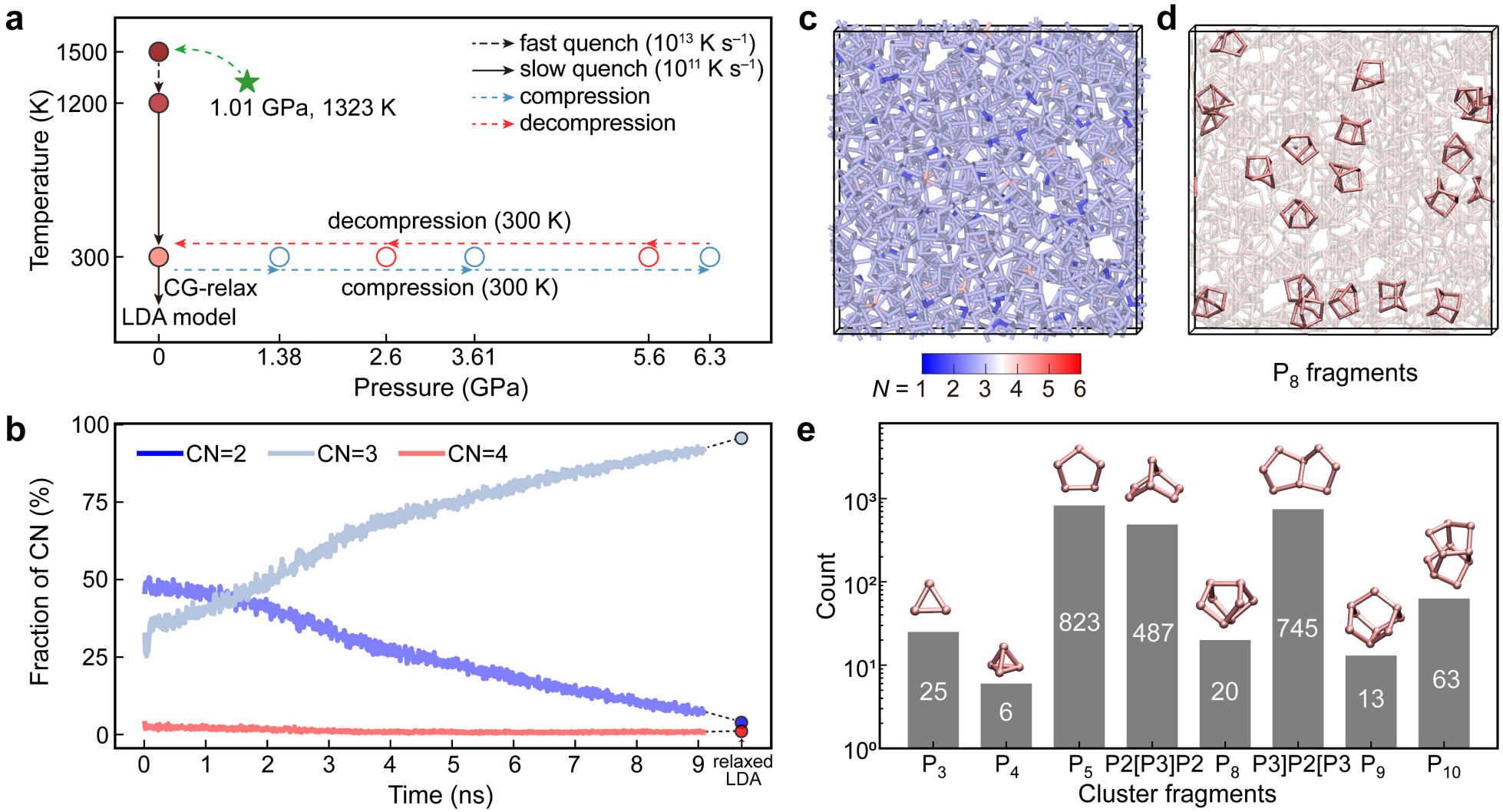
VLD, M. A. Caro, G. Csányi, *Nat. Commun.* **2020**, *11*, 5461

Towards automated & general ML potentials



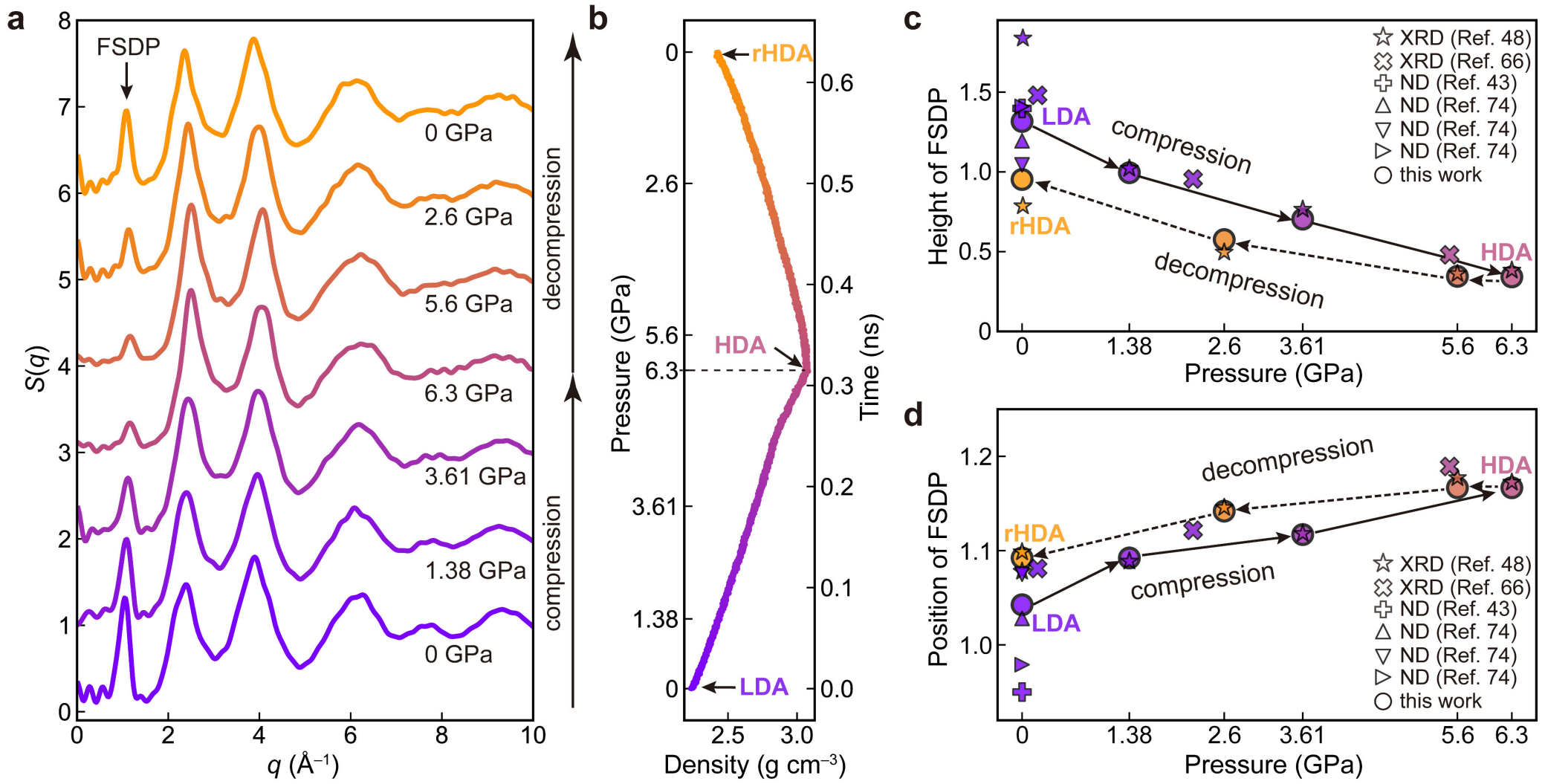
VLD, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073

The structure of amorphous red phosphorus



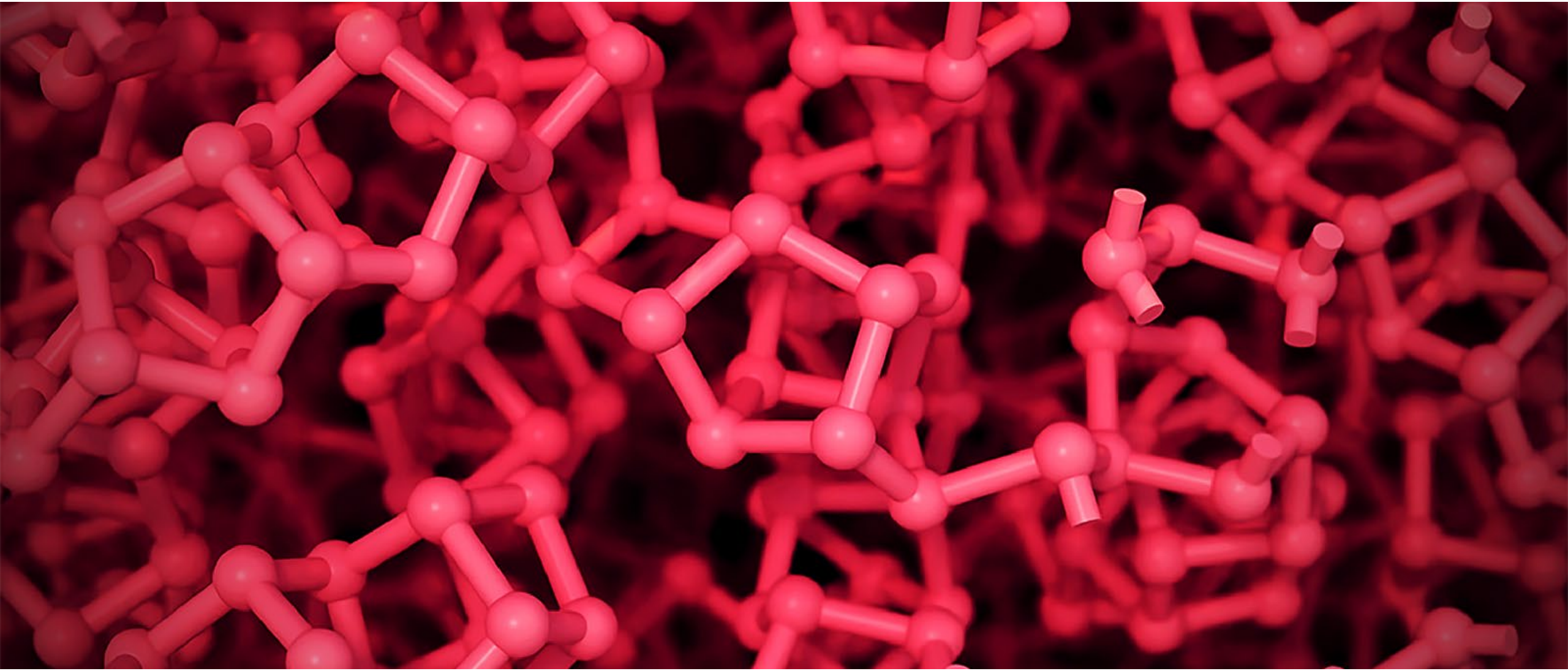
Y. Zhou, W. Kirkpatrick, VLD, *Adv. Mater.* **2022**, *34*, 2107515

The structure of amorphous red phosphorus



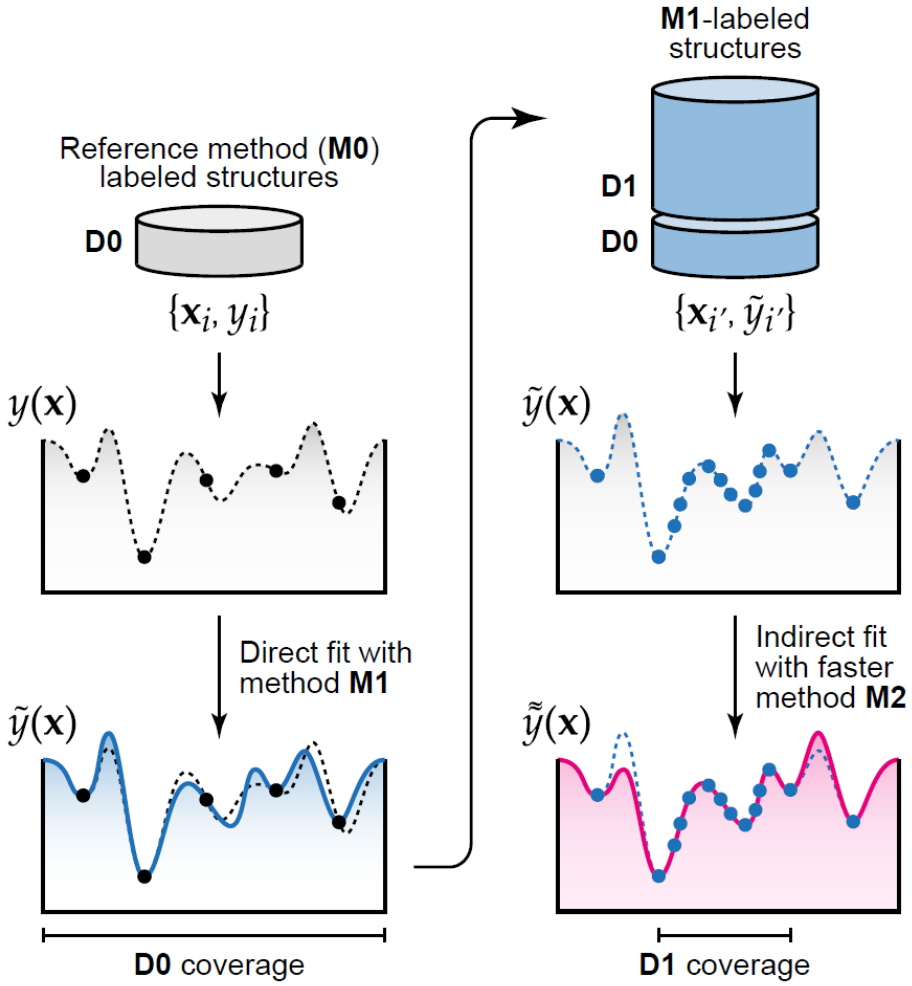
Y. Zhou, W. Kirkpatrick, VLD, *Adv. Mater.* **2022**, *34*, 2107515

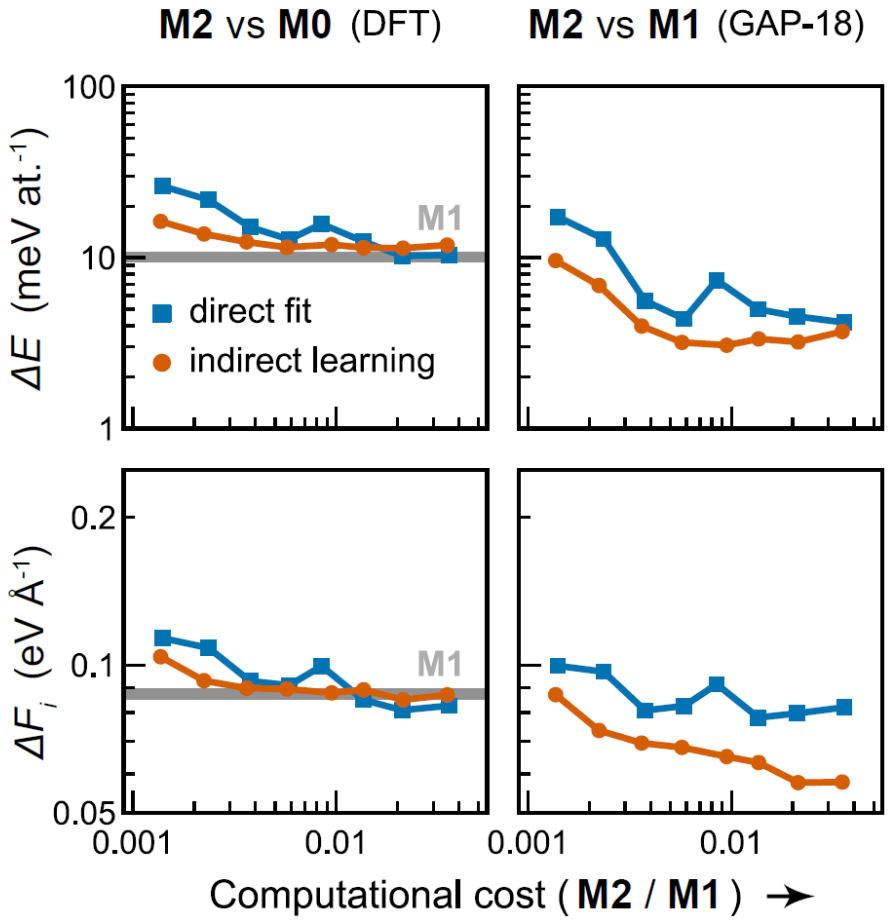
The structure of amorphous red phosphorus



Y. Zhou, W. Kirkpatrick, VLD, *Adv. Mater.* **2022**, *34*, 2107515

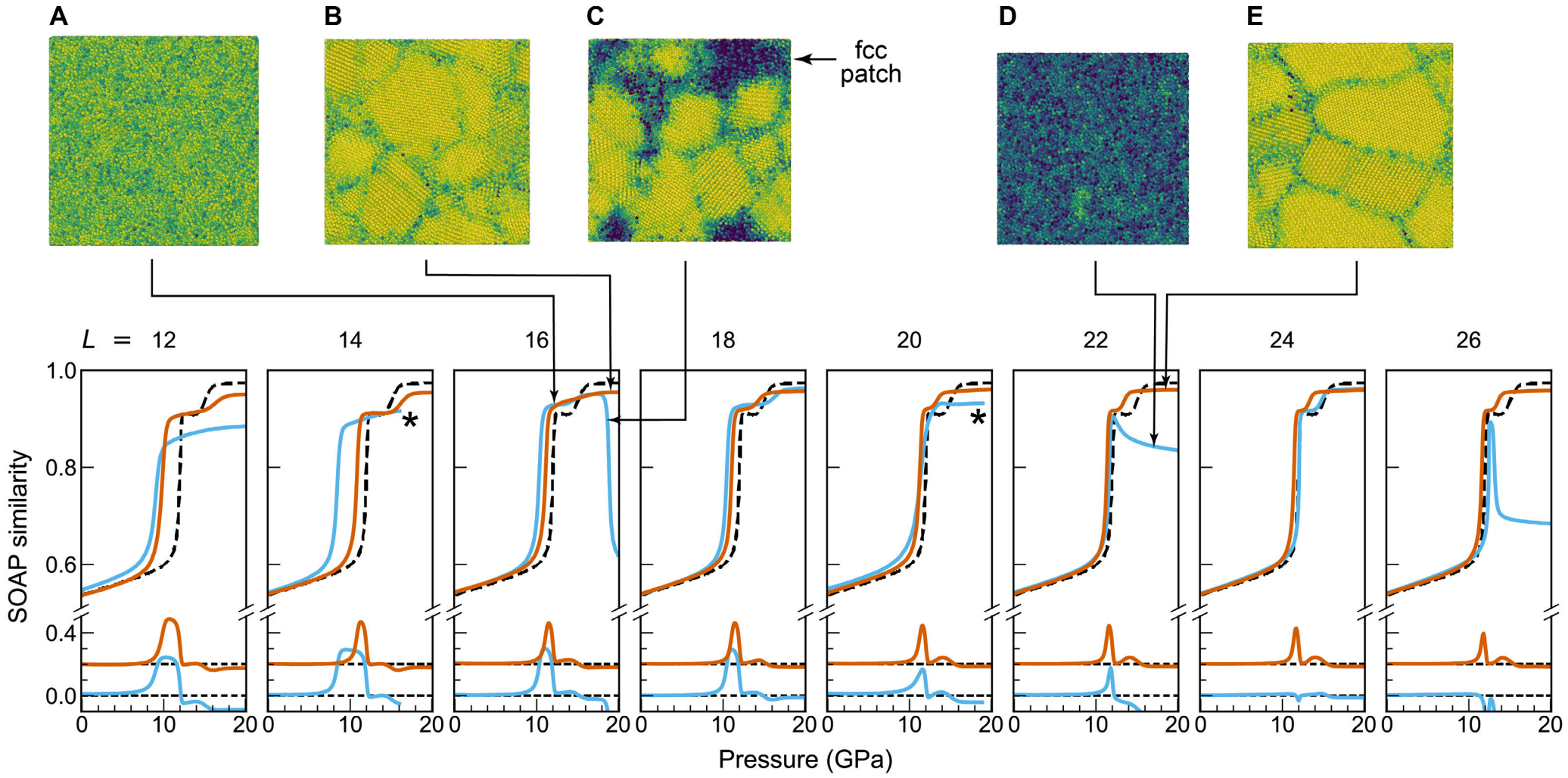
How to validate interatomic potentials





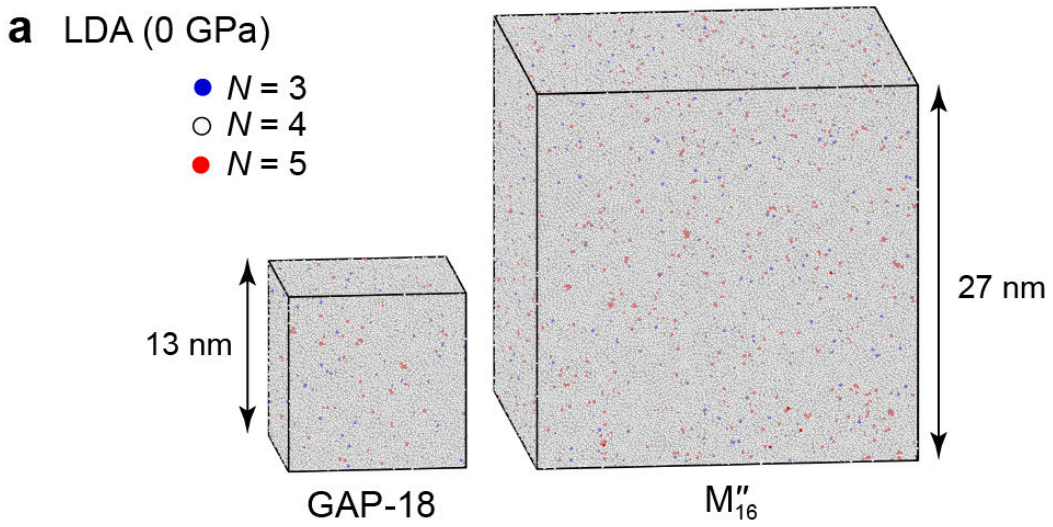
J. D. Morrow, VLD, *J. Chem. Phys.* **2022**, *157*, 104105

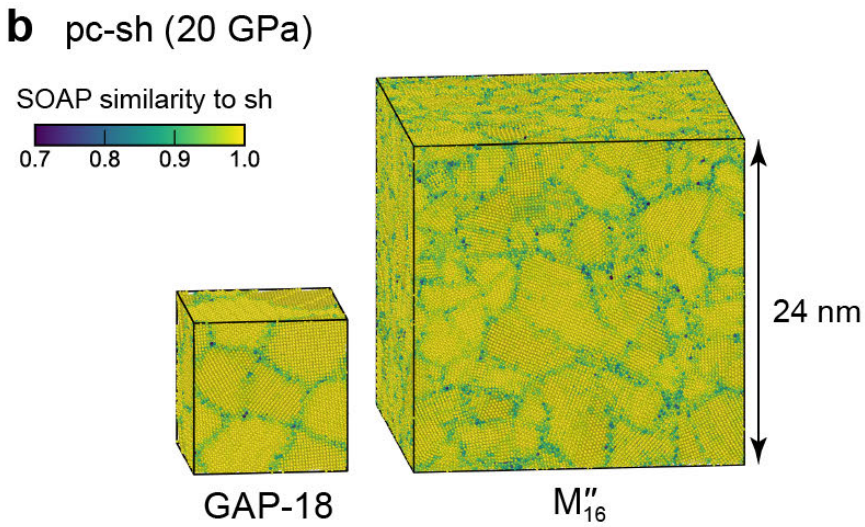
How to validate interatomic potentials

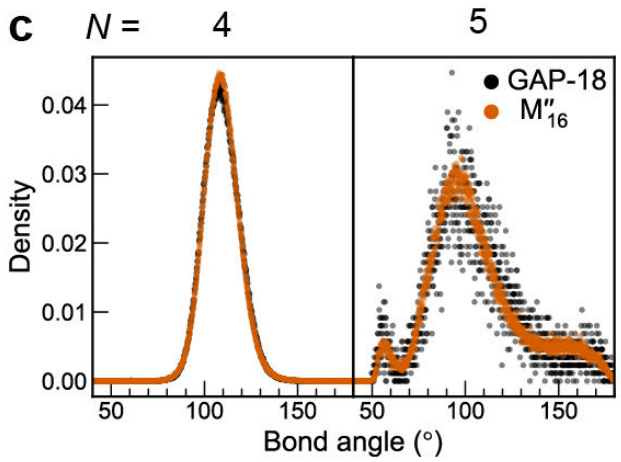


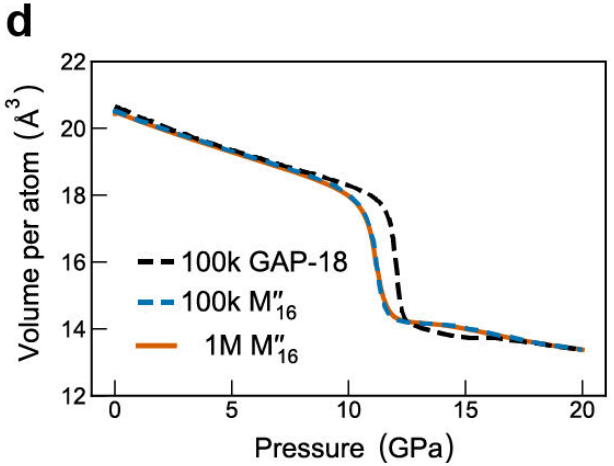
J. D. Morrow, VLD, *J. Chem. Phys.* **2022**, *157*, 104105

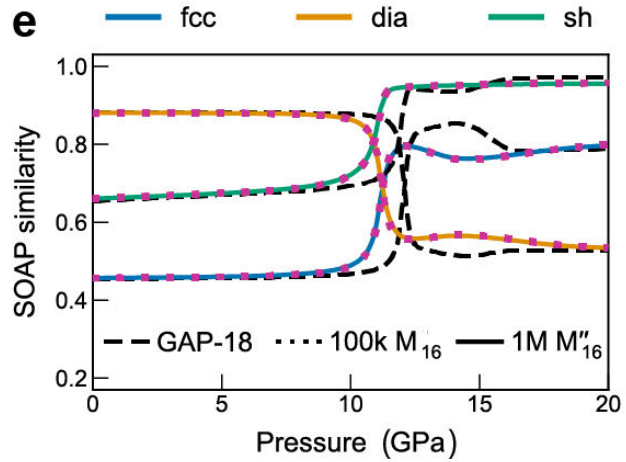
How to validate interatomic potentials











J. D. Morrow, VLD, *J. Chem. Phys.* **2022**, *157*, 104105

Executive summary

- **ML potentials** are increasingly popular simulation tools for materials modelling: accurate, flexible, and *fast*

  (*Adv. Mater.* **2019**, *31*, 1902765; *Chem. Rev.* **2021**, *121*, 10073)

- They are therefore giving new insight into **real materials**, from fundamental questions to full device-scale simulations

  (a-Si: *Nature* **2021**, *589*, 59; phase-change materials: arXiv:2207.14228)

- **Validating potentials** is becoming increasingly important, and many ideas here will be transferable from GAP to other models

  (arXiv:2211.12484 – see also tutorial by Joe)

Thank you!

Collaborators

ML potential development:
Gábor Csányi (Cambridge)

Amorphous silicon:
Noam Bernstein (US NRL)
Michele Ceriotti (EPFL)
David Drabold (Ohio)
Stephen Elliott (Cambridge)
Mark Wilson (Oxford)

Phase-change materials:
En Ma (Xi'an Jiaotong)
Wei Zhang (Xi'an Jiaotong)

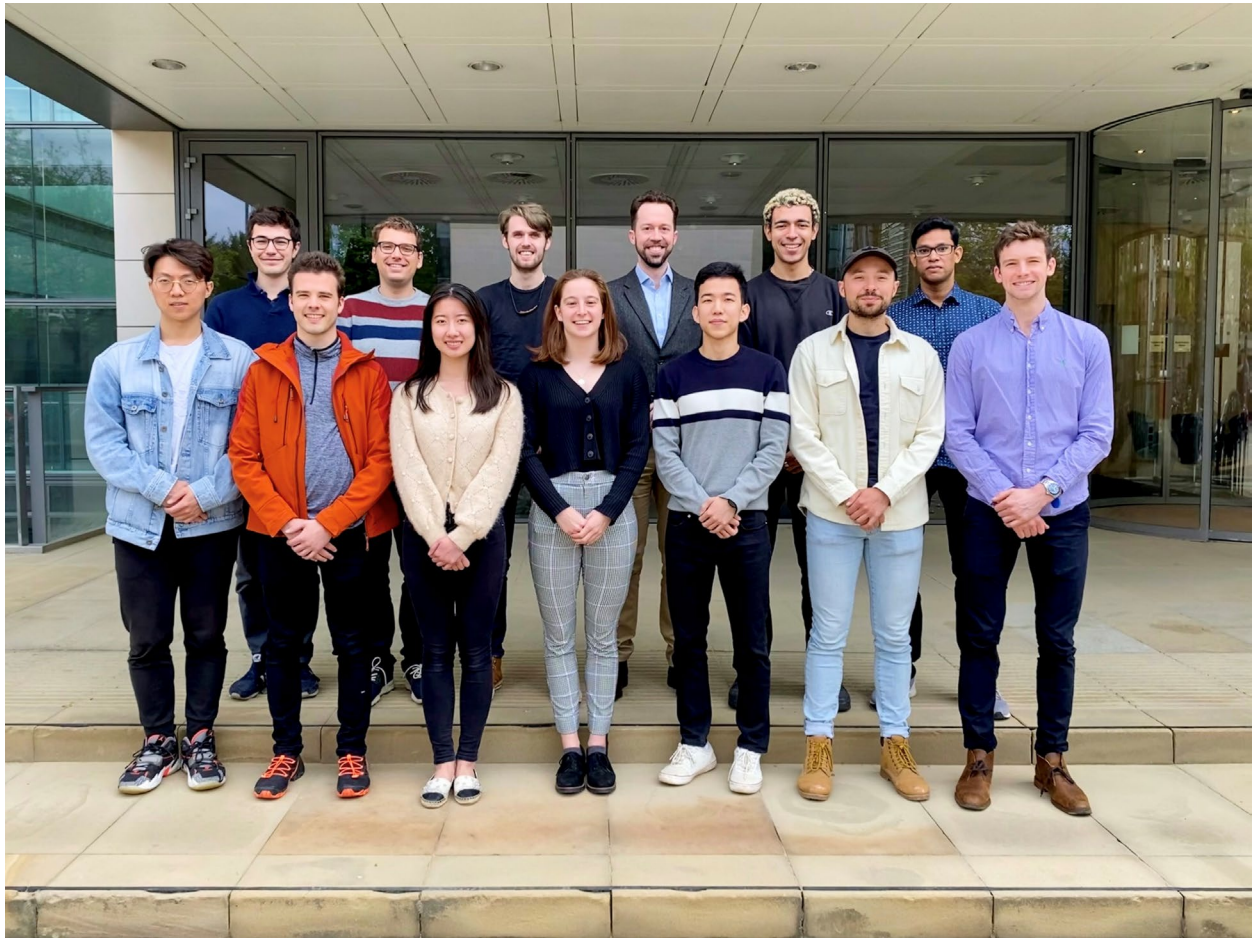
…and other key collaborators, including:
Karsten Albe (Darmstadt)
Miguel Caro (Aalto)
Janine George (BAM Berlin)
Andrew Goodwin (Oxford)
Davide Proserpio (Milan)

VLD Group Oxford

Research Associates
Dr Nijamudheen Abdulrahiman

DPhil (PhD) students
Zakariya El-Machachi
Zoé Faure Beaulieu
John Gardner
Joe Morrow
Tom Nicholas
Louise Rosset
Daniel Thomas du Toit
Yuxing Zhou

Part II (MChem) students
Kathryn Baker
Muxue Chen
Maximilian Weber
Aleksandra Zawadzka



UK Research and Innovation

(ERC guarantee funding)

Engineering and Physical Sciences Research Council

LEVERHULME TRUST

CHINA SCHOLARSHIP COUNCIL