Introduction to Material Modeling

2

Jan Hermann

Abstract

This introductory chapter presents material modeling, or computational materials science, as a discipline that attempts to computationally predict properties of materials and design materials with given properties, and discusses how machine learning can help in this task. At the center of material modeling is the relationship between the atomic structure and material properties. This relationship is introduced together with the quantum and statistical mechanics as the two main tools for material modeling from first principles, which should also guide the development of new machine learning approaches to computational materials science. Material modeling operates with an abundance of technical terms, and some of these are explained in a glossary at the end of the chapter.

2.1 Introduction

Material modeling, or computational materials science, refers to the problem of computational prediction of material properties, and also to the inverse problem of computational discovery of materials with given properties [1–4]. Traditionally, modeling of materials is based on the underlying fundamental physical laws—quantum and statistical mechanics—expressed in the form of mathematical equations, which provide an in principle exact framework for building the right models. This bottom-up approach is limited by the complexity of the equations, and by the difficulty in recognizing a priori which of the degrees of freedom in the

J. Hermann (✉)
TU Berlin, Machine Learning Group, Berlin, Germany

FU Berlin, Department of Mathematics and Computer Science, Berlin, Germany
e-mail: jan.hermann@fu-berlin.de

description of a system are relevant for the solution of the equations, and which can be omitted to reduce the complexity. Material modeling sometimes resorts to the opposite top-down approach as well, in which known materials and their properties are used to parametrize or simplify the models, but such techniques are often used in an ad-hoc fashion without any systematic theoretical guidelines. Here, machine learning as a discipline can fill this void and provide precisely such guidelines in the form of a solid mathematical framework for inferring unknown models from known data, extracting the relevant degrees of freedom along the way. Delivering on such a promise, this book presents a series of works that pave the way towards incorporating machine learning into material modeling. The challenge in such a program is to avoid creating an entirely orthogonal branch of research to the established techniques. Instead, the goal is to either incorporate machine learning into existing methods, thus enhancing them, or use the existing methods to bootstrap the machine learning approaches. For this reason, this introductory chapter gives a brief introduction into the traditional bottom-up approach of material modeling in the form of fundamental physical laws.

One possible way to characterize learning is that it involves inferring probability distributions from data that are assumed to be sampled from those distributions. In the common meaning of the word learning, the "data" may be observed instances of behavior and the probability distributions would be the "rules" according to which one behaves, both in terms of what behavior is possible (marginal distributions) and which behavior is appropriate for a given situation (conditional distributions). In some types of learning problems, with seemingly or evidently deterministic rules, the formulation in terms of probabilities may seem superfluous, but even in such cases it enables a rigorous handling of the fact that one always works with limited amount of data, and as such there can never be a certainty about the underlying inferred rules. In traditional machine learning tasks, e.g., image recognition, natural language processing, or optimal control, it is in general assumed that very little is known about the underlying distributions besides some general and often obvious properties such as locality or certain invariances. Either the data being modeled are too complex to derive anything meaningful about the distributions from theory, or, more commonly, there is no theory available.

When applying machine learning to molecules and materials, on the other hand, one has the whole toolbox of quantum and statistical mechanics to derive all sorts of general approximations or exact constraints to the distributions that underlie the data. The approximations can serve as a basis which a machine learning model improves, or as motivation for constructing specialized features and architectures. The constraints, such as symmetries or asymptotic behavior, can improve data efficiency and numerical stability of the learning algorithms, and in the case of unsupervised learning may even be a necessity to avoid learning unphysical solutions. In many applications, the methods of quantum and statistical mechanics do not enter the machine-learning techniques directly, but they are used to generate the data that are used to train the machine-learning models. In such cases, the question of feasibility and accuracy of the generated data becomes an integral part of the design of the machine learning approach.

This chapter aims to give a brief introduction to the problem of modeling of molecules and materials intended for a reader with only the most rudimentary knowledge of physics and chemistry. It attempts to provide a context for the technical parts of the book, and place the subsequent chapters in the broader map of materials science. The topic and scope of this chapter preclude any chances at being comprehensive, but it should give an explanation of some of the core principles and ideas, and point to topics of possible further study.

Section 2.2 presents the traditional problems encountered in material modeling in the framework of the relationship between material structure and material properties. Sections 2.3 and 2.4 present quantum mechanics and statistical mechanics—the two main disciplines of physics and theoretical chemistry that provide the theoretical framework for modeling molecules and materials. The chapter is concluded in Sect. 2.4 by a glossary of common terms that are frequently used in the subsequent technical chapters and where they are assumed to be understood.

2.2 Structure–Property Relationship

The ultimate goal of material modeling is to replace real-world lab experiments using materials and measuring instruments with (cheaper) computer simulations in the task of predicting the physical and chemical properties of a material [5]. This is the so-called *forward* problem in materials science. The corresponding *inverse* problem is to produce a (usually novel) material that has a set of desired properties, using computational techniques—this branch of materials science is called materials design or discovery [6]. The inverse problem is usually considered to be much harder than the forward problem, at least with existing techniques.

The reason for calling the material property prediction a forward problem is that, in principle, it has an exactly known unique solution. The mathematical solutions of the Schrödinger equation of quantum mechanics, applied to materials, can give answers to many questions about the properties of a material [7]. The rest of the questions can be, again in principle, answered by the tools of statistical mechanics [8, 9]. Unfortunately, the Schrödinger equation can be solved analytically only for the simplest "molecule"—a single hydrogen atom—and the numerical techniques that give sufficiently accurate solutions are computationally too expensive for many materials of interest. A large portion of the efforts in material modeling then deals with a careful analysis of these numerical models and their errors, with attempts to have these errors under control, and ultimately with devising more accurate, more general, and more computationally efficient models. The first technical part of the book, Representations, covers attempts to entirely avoid having to deal with the Schrödinger equation by learning the properties obtained from its solutions directly from data. Similarly, statistical mechanics can provide analytical answers only for the most rudimentary models, such as noninteracting harmonic oscillators, hard spheres, or two-dimensional lattices, and relies on statistical sampling from various types of thermodynamic distributions (*ensembles*) for more realistic systems. The biggest (unsolved) problem is then to generate representative samples from these

distributions. The second part of the book, Atomistic Simulations, deals with applications of machine learning to this sampling problem. The inverse problem of materials science, as every inverse problem, is likewise solvable in principle, by enumerating all possible materials and solving the forward problem for all of them. But this is unfeasible in most cases due to the sheer combinatorial size of the problem space in question. The third part of the book, Discovery, covers the use of machine learning to accelerate materials design by reducing the effective search space.

2.2.1 Atomic Structure

An important part of the discussion of material properties is the question of unique specification of a material or a molecule. All materials are composed of atoms, and specifying a material is usually understood as specifying its *atomic structure*. Disregarding quantum mechanics for a moment, the state of a piece of material is fully given by listing the positions and velocities of all its atoms, and to which element each of them belongs—this is called a *microstate*. However, since one gram of a typical solid or liquid contains on the order of 10^{21}–10^{23} atoms, this is practically impossible and in fact not necessary. The vast majority of the degrees of freedom involved in such a specification are constantly changing in any given piece of material, and as such do not pertain to any kind of permanent structure.

Which of the degrees of freedom constitute the atomic structure depends largely on the type of material and sometimes on the material property being investigated in the structure–property relationship. Although the term "material" is usually reserved for solid objects, consider water vapor for a moment as an example of a simple material. Water vapor consists of individual water molecules, each of which consists of one oxygen (O) and two hydrogen (H) atoms and whose geometry can be specified by the two O–H distances and the angle between them. At any given moment, the distances and angles in all the molecules in a drop of water are narrowly distributed around some mean values. (The actual width of these distributions increases with temperature and is a subject matter of statistical mechanics.) The same distributions would be found if one repeatedly measured the geometry of a single molecule over a period of time. The individual molecules move fast along straight lines between relatively rare collisions, spinning as they do, so the relative positions of the molecules change rapidly. It then follows that the only structurally persistent motif is the mean geometry of a water molecule—two O–H distances and one H–O–H angle. In the case of water vapor, this is all that can be said about its atomic structure, and many properties of water vapor can in fact be determined from studying just a single molecule as a representative of all the molecules.

Going to liquid water, the geometry of the individual molecules is unchanged, and they still constantly vibrate, move around, and spin, but they are now *condensed* and essentially touching each other. Many interesting water properties do not follow from the properties of a single molecule like in vapor anymore, but rather are a result

of the statistical characteristics of the relative positions and movement of the water molecules, which can be very complex. However, these characteristics do uniquely follow from the interactions between individual water molecules. In principle, there could be different "kinds" of water, for which the statistical characteristics and hence material properties would be different, but this is not the case, and neither is for vast majority of liquids. As a result, all it takes to specify liquid water as a material is the geometry of a water molecule.

Moving on to solid water—ice—the water molecules themselves are again unchanged, they are condensed similarly to liquid water, and they still vibrate around the most likely geometry, but now their relative positions and rotations are frozen. As in the case of the *intra*molecular degrees of freedom (O–H distances, H–O–H angles), the *inter*molecular degrees of freedom are not sharp values, but narrowly distributed around some mean values, nevertheless they are not entirely free as in the liquid or vapor. Fortunately, it is not necessary to specify all the relative positions of all the molecules in a piece of ice. Common ice is a *crystal*, which means that its water molecules are periodically arranged with a relatively short period. Thus one needs to specify only the positions of several water molecules and the periodic pattern to express the atomic structure of ice. Unlike in the case of liquid water, the water molecules in ice can be arranged in different structural arrangements, and as a result there are different types (*phases*) of ice, each with different material properties. (Almost all ice found naturally on Earth is of one form, called I_h.) In general, a crystal is any material with periodic atomic structure. This relative structural simplicity enables them to be specified and characterized precisely, because a microscopic region of a crystal at most several nanometers across uniquely determines the atomic structure of the whole macroscopic piece of a material. Many common solid materials are crystalline, including metals, rocks, or ceramics. Usually a macroscopic piece of a material is not a single crystal, but a large array of very small crystals that stick together in some way. This structure on top of the microscopic atomic structure of a single crystal can sometimes have an effect on some properties of a material, but often is inconsequential and not relevant for material modeling.

Other solid materials, including glasses, plastics, and virtually all biological matter, are not crystalline but amorphous—their atomic structure has no periodic order. Still, specifying the structure of such materials does not require specifying all the atomic positions. Consider a protein as an instance of complex biological matter. Proteins are very large molecules consisting of hundreds to hundreds of thousands atoms, with typical sizes in thousands of atoms, that play a central role in all life on Earth. All proteins are constructed as linear chains, where each chain link is one of 22 amino acids, which are small organic molecules. Like in a common chain, the overall three-dimensional shape of a protein is relatively easy to change, whereas the linear sequence of the links (amino acids) is fixed. Unlike in a common chain, the amino acids themselves have internal degrees of freedom, with some of them hard as the distances and angles in the water molecules, and some of them relatively soft. When a protein molecule is constructed in a cell from the amino acids, its three-dimensional shape is essentially undetermined, and all that can be said about

its atomic structure in that moment is the linear sequence of the amino acids—this is called the *primary* structure. Most proteins in living organisms are not found having random shapes, but very specific *native* forms. As a result of the complex quantum-mechanical interactions within the protein and of the protein with its environment, the initially random shape *folds* into the native shape, referred to as the secondary and tertiary structure, at which point the protein can fulfill its biological function. This process does not only fix the overall three-dimensional shape of the protein, but also some of the previously free internal degrees of freedom within the amino acids. In the native state, the atomic structure of a protein can be characterized much more precisely. Finally, many proteins can be extracted from the cells into solutions and then crystallized, which freezes all the degrees of freedom, just like when ice is formed from water. At that point the atomic structure can be specified most precisely, with all atoms of the protein vibrating only slightly around their mean positions.

The previous examples demonstrate that what is considered to be the atomic structure of a material depends on a given context. The myriads of degrees of freedom in the atomic structure of a piece of a material can be under given circumstances (temperature, pressure, time scale) divided into hard ones and soft ones, with the caveat that there is no sharp boundary between the two. The soft degrees of freedom are described by distributions with large variance, sometimes multi-modal, and often with strong correlations between them. These distributions follow the laws of statistical mechanics, and in many cases are detrimental to the molecular and material properties, but are not considered to be part of the atomic structure per se. The hard degrees of freedom can be described by sharp probability distributions with small variance, and their mean values constitute the atomic structure of a molecule or a material. The chapters of the first part of the book deal with how best to represent the atomic structure so that machine learning approaches can then efficiently learn the relationship between the structure and the properties. These constitute both properties that follow directly from the structure and are independent of the soft degrees of freedom, as well as those that follow from the statistics of the soft degrees of freedom, which are in turn constrained by the hard degrees of freedom. The following section discusses the material properties in more detail.

2.2.2 Molecular and Material Properties

Having clarified what is meant by the structure of a molecule or a material, the two problems of materials science can be stated as being able to maximize the following two conditional probability distributions:

$$\text{forward:} \qquad P(\text{property}|\text{structure})$$

$$\text{inverse (materials design):} \qquad P(\text{structure}|\text{property})$$

The atomic structure was introduced above as a specification of a material that uniquely determines its properties, which would suggest that the forward problem should be formulated in terms of a function, rather than a distribution. Nevertheless, it is useful to replace the deterministic function with a probability distribution for two reasons. First, the forward problem is strictly deterministic only under idealized conditions (e.g., isolated molecule, zero temperature) and when the problem is set up in such a way that the atomic structure is indeed specified fully. For instance, temperature effects or random defects in crystals smear many sharp properties into Gaussian distributions. Furthermore, a given material modeling problem may specify atomic structure only partially, such as when one is given only the structure of a molecule, of which a molecular crystal is composed, but not the complete structure—the periodic arrangement of the molecules. Second, when the forward problem is to be learned from data, rather than derived from first principles, the uncertainty, and hence probability, arises naturally from the finite amount of available data. In contrast to the forward problem, its inverse is inherently probabilistic, because molecular and material properties are never guaranteed to uniquely specify a molecule or a material.

Listing all the material properties that scientists have ever measured would cover many books on its own, so this introduction will keep to classifying them into general categories. The molecular and material properties can be tentatively divided into two categories: electronic properties on the one hand and thermodynamic properties on the other.

Atoms themselves are not indivisible, but consist of heavy nuclei (which can be almost always considered point-like in materials science) and thousand-times lighter electrons, which form electronic "clouds" around the nuclei. The behavior of the electrons deviates strongly from classical mechanics, and is best described by quantum mechanics. The closest (but still bad) classical analogy for the electrons would be perhaps that of a liquid floating around the nuclei, rather than particles moving around them. In any case, because most of the mass of an atom is concentrated in its nucleus, its position is likewise associated with the position of the nucleus. Under almost all circumstances, the atoms moving around can be described as nuclei moving around and being instantly followed by their respective clouds of electrons. The mathematical formulation of this idea is called the Born–Oppenheimer approximation, which underlies most of material modeling. The hard degrees of freedom that constitute the atomic structure are then associated with fixed electronic clouds, and the material properties that can be explained by quantum mechanics from these stationary electronic clouds are considered to be electronic properties. In contrast, the thermodynamic properties can be explained by statistical mechanics from the often complex statistics of the soft degrees of freedom in the motion of the atoms (atomic nuclei).

Electronic properties of molecules and materials are essentially properties of the collection of their electrons (electronic cloud) under the influence of the nuclei located at positions given by the atomic structure. As will be discussed in a bit more detail in Sect. 2.3, the electronic cloud can be in different discrete quantum states, and under temperatures found on Earth, most matter is found in the lowest-energy

state, called the *ground* state. The higher-energy states, called *excited* states, can be reached, for instance, by exposing a material to visible light or UV radiation. In general, one can then distinguish between the ground-state electronic properties and the excited-state properties, the latter usually referring to the *transitions* between the ground and the excited states. Some of the ground-state properties include the atomization energy (energy released when a molecule is formed from its constituent atoms), the dipole moment and polarizability (shape and responsiveness of the electronic cloud), and the vibrational spectra (the hardness of the hard degrees of freedom). (See the Glossary for more details.) It is mostly the ground-state electronic properties that are targeted by the approaches in the first part of this book. The excited-state properties include UV and optical absorption spectra (which determine the color and general photosensitivity of a material), or the ability to conduct electrical current.

Thermodynamic properties stem from the motion of the atoms along the soft degrees of freedom in a material. Virtually all the soft degrees of freedom are not entirely free, but are associated with some barriers. For instance, in liquid water the molecules can rotate and move around, but not equally easily in all directions, depending on a particular configuration of the neighboring molecules at any given moment. The ability to overcome these barriers is directly related to the average speed of the atoms, which in turn is expressed by the temperature of the material. At absolute zero, $0\,\text{K}$, the atoms cease almost all motion, only slightly vibrating around their mean positions as dictated by quantum mechanics. At that point, all the degrees of freedom in the material can be considered hard, and virtually all material properties that can be observed and defined at absolute zero can be considered electronic properties. Calculations of material properties that neglect the motion of atoms essentially model the materials as if they were at zero temperature. As the temperature is increased, the atoms move faster and fluctuate farther from their mean positions, and the lowest barriers of the least hard degrees of freedom can be overcome, which turns them into soft degrees of freedom. The statistical characteristics of the resulting atomic motion, and material properties that follow from it, can be then obtained with the tools of statistical mechanics. (Statistical mechanics is valid and applicable at all temperatures, but only at higher temperatures does the atomic motion become appreciably complex.)

The study of thermodynamic properties—thermodynamics—can be divided to two main branches, equilibrium and nonequilibrium thermodynamics. The equilibrium refers to a state of a material in which the hard degrees of freedom and the probability distributions of the soft degrees of freedom remain unchanged over time. Traditionally, thermodynamic properties refer to those properties related to the motion of atoms that can be defined and studied at equilibrium. This includes the melting and boiling temperatures, the dependence of density on pressure and temperature, the ability to conduct and absorb heat, or the ability of liquids to dissolve solid materials. It is mostly properties of this kind that motivate the development of the approaches presented in the second part of this book. On the other hand, nonequilibrium thermodynamics deals with *processes* that occur in materials when they are out of equilibrium, that is, when the distributions of the

degrees of freedom change over time. Examples of such processes are chemical reactions and their rates, transport of matter through membranes, or the already mentioned folding of a protein to its native state.

2.3   Quantum Mechanics

Quantum mechanics is the set of fundamental laws according to which all objects move [10]. In the limit of macroscopic objects, with which we interact in everyday life, the laws of quantum mechanics reduce to classical mechanics, which was first established by Newton, and for which all of us build good intuition when jumping from a tree, riding on a bus, or playing billiard. On the microscopic scale of atoms, however, the laws of quantum mechanics result in a behavior very different from that of the classical mechanics.

One of the fundamental differences between the two is that in quantum mechanics, an object can be simultaneously in multiple states. In classical mechanics, an object can be exclusively either at point $\mathbf{r}_1$ or at point $\mathbf{r}_2 \neq \mathbf{r}_2$, but not in both. In quantum mechanics, an object, say an electron, can be in any particular combination of the two states. Mathematically, this is conveniently expressed by considering the two position states, denoted $|\mathbf{r}_1\rangle$, $|\mathbf{r}_2\rangle$, to be basis vectors in a vector space (more precisely a Hilbert space), and allowing the object to be in a state formed as a linear combination of the two basis vectors,

$$|\psi\rangle := c_1|\mathbf{r}_1\rangle + c_2|\mathbf{r}_1\rangle$$

Note that $c_1|\mathbf{r}_1\rangle + c_2|\mathbf{r}_1\rangle \neq |c_1\mathbf{r}_1 + c_2\mathbf{r}_2\rangle$, that is, the object is not simply in a position obtained by adding the two position vectors together, but rather it is somewhat at $\mathbf{r}_1$ and somewhat at $\mathbf{r}_2$. Generalizing the two position vectors to the infinite number of positions in a three-dimensional space, the general state of an object can be expressed as

$$|\psi\rangle := \int d\mathbf{r}\, \psi(\mathbf{r})\, |\mathbf{r}\rangle$$

where $\psi(\mathbf{r})$, called a *wave function* plays the role of the linear coefficients c_1, c_2 above. It is for this reason that electrons in molecules are better to be thought of as electronic clouds (corresponding to $\psi(\mathbf{r})$) rather than point-like particles.

Another fundamental law of quantum mechanics is that each in principle experimentally observable physical quantity ("observable" for short) is associated with a linear operator, $\hat{L}$, acting on the Hilbert space of the object states, and the values of the quantity that can be actually measured are given by the eigenvalues, λ_i, of the operator,

$$\hat{L}|\psi_i\rangle = \lambda_i|\psi_i\rangle$$

One of the most important operators is the energy operator, called Hamiltonian, which determines which energies of an object can be measured,

$$\hat{H}|\psi_i\rangle = E_i|\psi_i\rangle$$

This particular eigenvalue equation is called a Schrödinger equation, and since energy plays a key role in all physical phenomena, its solution (the eigenvalues and eigenstates) enables determination of many electronic properties of both the ground state ($|\psi_0\rangle$) and the excited states ($|\psi_n\rangle$, $n > 0$). The abstract eigenvalue equation can be transformed into a differential equation by expressing the state vectors using the wave function,

$$\hat{H}\psi_i(\mathbf{r}) = E_i\psi_i(\mathbf{r})$$

where $\hat{H}$ becomes a differential operator. For example, consider a hydrogen atom, consisting of a single electron described by position $\mathbf{r}$ and moving around a nucleus fixed at position $\mathbf{R}$. Such a system can be considered the simplest "molecule." In this case, the Schrödinger equation for the wave function of the electron (in atomic units) has the form

$$\left(-\frac{1}{2}\nabla^2 - \frac{1}{|\mathbf{r} - \mathbf{R}|}\right)\psi_i(\mathbf{r}) = E_i\psi_i(\mathbf{r})$$

where ∇^2 is the Laplace operator. In more complex molecules with multiple nuclei and electrons, the Hamiltonian contains more terms, and the wave function is a function of the coordinates of all the electrons, but the general structure of the problem remains the same.

The Schrödinger equation for the hydrogen atom can be solved analytically ($E_0 = -\frac{1}{2}$, $\psi_0(\mathbf{r}) \propto e^{-|\mathbf{r}-\mathbf{R}|}$), but this is not possible for more complex molecules. Direct methods for numerical solution of differential equations are inapplicable because of the dimensionality of the wave function ($3N$ dimensions for N electrons). Many methods of quantum chemistry, such as the Hartree–Fock or the coupled-cluster method, attempt to solve this issue by cleverly selecting a subspace in the full Hilbert space of the electrons spanned by a finite basis, and finding the best possible approximate eigenstates within that subspace. This turns the differential Schrödinger equation into an algebraic problem, which can be solved numerically at a feasible computational cost. Another class of methods, such as the density functional theory, also change the original Hamiltonian such that the eigenvalues (and in some cases also the eigenstates) are as close to the true eigenvalues as possible. In all the approximate quantum-mechanical methods, however, the computational cost grows asymptotically at least as $O(N^3)$ (or much faster for the more accurate methods), and their use becomes unfeasible from a certain system size. These include the density functional theory ($O(N^3)$), the Hartree–Fock (HF) method ($O(N^3)$), the Møller–Plesset perturbation theory to second order (MP2, $O(N^3)$), the coupled-cluster method with single, double,

and perturbative triple excitations (CCSD(T), $O(N^7)$), and the full configuration interaction (FCI, $O(\exp(N))$).

Once the Schrödinger equation is solved, evaluating the electronic properties from the known eigenvalues and eigenstates of the Hamiltonian is often straightforward. In particular, many properties such as the dipole moment, polarizability, or atomic forces can be calculated as integrals of the corresponding operators over a given eigenstate, or as derivatives of such integrals, which can be transformed to integrals over derivatives of the operators by the Hellmann–Feynman theorem [11]. Besides that, the solution to the Schrödinger equation also provides a direct link between the quantum mechanics of the electrons and the statistical mechanics of the atoms. The electronic Hamiltonian has terms depending on the positions of the nuclei, $\mathbf{R}_i$, and as a result, the energies of the eigenstates likewise depend on the positions of the nuclei. The energy of a particular eigenstate as a function of the nuclear positions, $V(\mathbf{R}_1, \ldots)$, is called a *potential energy surface*, and in principle completely determines the dynamics of the motion of the atoms.

2.4 Statistical Mechanics

In the context of material modeling, statistical mechanics deals with the motion in the soft degrees of freedom of the atoms in a material, and its central idea is that most of the detail in that motion ($> 10^{21}$ variables) can be safely omitted, while the physically relevant characteristics can be expressed in a smaller number of collective degrees of freedom. These collective variables can range from microscopic, in the form of a coarse-grained description of an atomistic modeling, to macroscopic, such as temperature or pressure. In all cases, the remaining degrees of freedom beyond the collective ones are treated in a statistical fashion, rather than explicitly.

The fundamental concept in statistical mechanics is that of a microstate, $\mathbf{s}$, which comprises the positions, $\mathbf{R}_i$, and velocities $\mathbf{v}_i$, of all the atoms in a material, $\mathbf{s} \equiv (\mathbf{R}_1, \mathbf{v}_1, \mathbf{R}_2, \ldots) \equiv (\mathbf{v}, \mathbf{R})$. The total energy, H, of a given microstate consists of the kinetic part, arising from the velocities, and the potential part, which is determined by the potential energy surface,

$$E(\mathbf{s}) = \sum_i \frac{1}{2} m_i v_i^2 + V(\mathbf{R}_1, \ldots)$$

One of the central results of statistical mechanics is that given that a material is kept at a constant temperature T (the so-called canonical ensemble), the probability density of finding it at any particular microstate is proportional to the so-called Boltzmann factor (in atomic units),

$$P(\mathbf{s}) \propto \mathrm{e}^{-\frac{E(\mathbf{s})}{T}} \quad \Rightarrow \quad P(\mathbf{R}) \propto \mathrm{e}^{-\frac{V(\mathbf{R})}{T}}$$

The latter proportionality follows from the fact that the kinetic and potential parts of the total energy are independent. Close to absolute zero temperature, the Boltzmann factor is very small (in relative sense) for all but the lowest-energy microstates, which correspond to small atomic velocities and atomic positions close to the minimum of the potential energy surface. This coincides with the picture of all the degrees of freedom in the atomic motion being hard close to absolute zero. As the temperature rises, the microstates with higher energy become more likely, corresponding to higher velocities and atomic positions further from the energy minimum. To see how this simple principle can be used to calculate thermodynamic properties of materials, assume that one can enumerate all the possible microstates, calculate the sum of all the Boltzmann factors—the *partition function* [12]—and thus normalize the probabilities above,

$$Z(T) = \int d\mathbf{s}\, e^{\frac{-E(\mathbf{s})}{T}}, \qquad P(\mathbf{s}) = \frac{1}{Z(T)} e^{-\frac{E(\mathbf{s})}{T}}$$

The mean total energy, for instance, can be then calculated directly from the partition function,

$$\langle E \rangle = \int d\mathbf{s}\, P(\mathbf{s}) E(\mathbf{s}) = T^2 \frac{\partial \ln Z}{\partial T}$$

A quantity closely related to the partition function is the *free energy*,

$$F(T) = T \ln Z(T)$$

Whereas the total partition function of two combined systems is a product of their respective partition functions, the logarithm in the definition of the free energy makes it an additive quantity. The multiplication by the temperature allows the free energy to have a physical interpretation—the change in the free energy between two states of a system is the maximum amount of work that can be extracted from a process (is available—free) that takes the system from one state to the other.

One of the reasons that makes the free energy (and partition function) a powerful tool is that it can be calculated not only for the set of all possible microstates, but also for physically meaningful subsets. For instance, consider the melting temperature of a solid as an example. To use the free energy, one can characterize the melting temperature as the point of inversion of the probabilities of the atoms of a material appearing solid on the one hand or liquid on the other (inversion of free energies of a solid and of a liquid),

$$p_{\text{solid}}(T) \propto \int_{\text{solid}} d\mathbf{s}\, e^{-\frac{E(\mathbf{s})}{T}} = Z_{\text{solid}}(T) = e^{-\frac{F_{\text{solid}}(T)}{T}}, \qquad p_{\text{liquid}}(T) \propto e^{-\frac{F_{\text{liquid}}(T)}{T}}$$

where the integrals run over all microstates that correspond to the solid or liquid forms of matter. The melting temperature can then be calculated as the temperature

at which the free energies of the solid and liquid forms of a given material are equal (the probabilities of finding the two forms are identical). In the case of melting, the average potential energy of the solid microstates is lower (Boltzmann factors larger) than that of the liquid microstates, but there is many more liquid microstates than solid microstates, so the free energies (total probabilities) end up being equal.

The computational difficulty in statistical mechanics lies in the evaluation of the integrals over microstates such as those above. Even if one modeled only several hundreds atoms of a material at a time, the number of degrees of freedom involved in a microstate prohibits analytical evaluation of the integrals, and even direct numerical integration is unfeasible. Fortunately, Monte Carlo techniques present a general approach to evaluate such high-dimensional integrals by replacing them with sums over representative samples of the microstates. The task is then to generate statistically significant and diverse microstates, that is, microstates with large Boltzmann factors that completely span the physically relevant microstate subspaces. This can be achieved by many different techniques, the most common one being *molecular dynamics*. In this approach, the atoms are let to move according to the laws of classical mechanics along trajectories influenced by the potential energy surface, and the microstate samples are obtained by taking periodic snapshots of this dynamical system. This approach is justified by the so-called *ergodic principle*, which states that integral averages over the space of microstates (also called the *phase space*) are equal to time averages over sufficiently long times when a system is let to evolve under appropriately chosen dynamical conditions. This technique is used in most chapters in the second part of the book.

Glossary

Atomic units   The standard units of measurement (SI units), including the meter, kilogram, or joule, are convenient for macroscopic settings, but result in very small values of quantities in the microscopic world. The atomic units have been designed to alleviate this inconvenience, and also to simplify physical equations by making the numerical values of common physical constants equal to one (in atomic units). For instance, the atomic unit of energy is called Hartree, and the ground-state electronic energy of a hydrogen atom is $-\frac{1}{2}$ Hartree. In SI units, this would be equal to approximately $-2.18 \cdot 10^{-18}$ J.

Boltzmann distribution   A piece of material in equilibrium at temperature T can be found in a microstate $\mathbf{s}$ (positions and velocities of all atoms) with a probability that is proportional to $e^{-H(\mathbf{s})/k_B T}$. This probability distribution of the microstates is called a Boltzmann distribution. Since the energy of a microstate consists of two independent parts—the kinetic and potential energy—the distribution of the positions of the atoms also follows a Boltzmann distribution. The statistical ensemble of microstates following the Boltzmann distribution is called a canonical ensemble.

Chemical bond   All chemistry is a consequence of the motion of electrons in molecules, which is in general complicated. Comparing the electronic motion across different molecules reveals common patterns, and chemical bonding is one of the most widely recognized of such patterns. For instance, when two carbon atoms get close to each other, between one to three pairs of electrons tend to concentrate between the two atoms, depending on other atoms in the neighborhood. This in turn attracts the two atoms together. This effect is an example of a chemical bond and is what holds the atoms in diamond together.

Computational cost   Usually called computational complexity in computer science, the computational cost of approximate methods for solving the fundamental physical equation of quantum mechanics and statistical mechanics involved in material modeling is one of their three key properties besides accuracy and universality. The main determining factor of a given factor is usually the size of the system being modeled, that is, the number of atoms.

Configuration vs. conformation   The degrees of freedom in the positions of atoms in a material can be under given circumstances, such as temperature, divided into hard and soft. The hard degrees are essentially fixed for the purpose of a given modeling task and determine the *configuration*. The soft degrees of freedom can change through the simulation and their particular arrangement is called a *conformation*. For instance, the sequence of amino acids in a protein is a configuration, whereas any particular three-dimensional shape of the amino acid chain is a conformation.

Crystal   The atomic structure of many solids is characterized by a small pattern of atoms (usually units to hundreds) periodically repeated throughout the three-dimensional space. Such a material is called a crystal. Most crystalline materials do not consist of single large crystals, but of many small crystals randomly stitched together, or forming a powder. Although crystals are usually modeled as being perfectly periodic, real-world crystals have various defects that make the perfect crystal only an approximation.

Density functional theory (DFT)   Tracking all the interactions and correlations in the motion of electrons in molecules and materials becomes quickly unfeasible as the system size grows. DFT attempts to alleviate this problem by reformulating the electronic problem such that the electrons do not interact explicitly one with each other, but rather only via the total density of electrons, which makes the problem mathematically tractable. DFT is in principle an exact theory, but its practical realizations (the different functionals of the electron density) achieve only approximate description of the electronic motion.

Dipole moment   When an electrical charge is distributed continuously in space, such as in the electronic cloud of molecules, its dipole moment is simply the mathematical first moment of the density of the charge, $\mathbf{p} = \int d\mathbf{r}\, \mathbf{r} n(\mathbf{r})$. Its importance lies in the fact that when two molecules are sufficiently far apart, their electric interaction can be Taylor expanded around the infinite separation, and the leading term of this expansion depends on the dipole moments of the molecules.

Electronic property   Material property that can be explained by the electronic structure of a material without regard for the statistics of the motion of the atoms.

Electronic structure   The electronic structure of a molecule or a material refers to the particular arrangement of electrons in it and their collective properties. The electronic structure can be obtained by solving the Schrödinger equation of quantum mechanics, and used to predict and explain electronic properties.

Excitation energy   The electrons in a molecule or a crystal can be collectively in different states with different electronic properties. Most matter on Earth is in the lowest-energy state, called the ground state, but electrons can be *excited* to higher-energy states using light or chemical reactions, which supply the necessary excitation energy. The excited states usually do not persist for long and fall back to the ground state, with the excitation energy being released back in various forms.

First principles (ab initio)   Approximate methods of material modeling that are considered to be based on first principles can be straightforwardly derived from fundamental physical laws without introducing much room for tuning. They stand in opposition to empirical approaches which use flexible models that can be optimized to reproduce available data. The first principles and empirical methods are not two binary categories, but opposite extremes on a spectrum.

Force field   The electronic structure methods that are able to calculate the true electronic energy for a given position of nuclei are usually too costly for them to be used for molecular dynamics simulations, which model often large systems and where the energy must be evaluated many times. Fortunately, the absolute value of the electronic energy is irrelevant for the molecular dynamics, only the forces exerted on the atoms matter. Force fields are usually relatively simple sets of functions that take the positions of atoms as an input, and map them to the forces acting on the atoms. Force fields are usually highly empirical and their parametrization from data is a tedious task.

Free energy   Free energy is always associated with some subset of microstates and is a direct measure of a probability, p, to find a system in that subset, $F = -T \ln p$. The subsets are usually either coarse-grained degrees of freedom (e.g., all microstates of a protein with a given overall three-dimensional shape fixed, regardless of the internal degrees of freedom of individual amino acids) or macroscopic states (e.g., all microstates with a given total volume of the system). Only differences between the free energies of different microstate subsets are physically relevant.

HOMO–LUMO gap   The difference between the energies of the highest-occupied molecular orbital (HOMO) and the lowest-unoccupied molecular orbital (LUMO). This quantity can only be defined within approximate models of the electrons in molecules and is not physically observable, but it is an approximation of the lowest excitation energy.

Hamiltonian   The Hamiltonian is a physicist's way of uniquely specifying a given physical system, by relating the energy of a system to its internal degrees of freedom. Given a Hamiltonian, the behavior of a system can be in principle calculated using the laws of quantum mechanics, or approximately by classical

mechanics. In quantum mechanics, the Hamiltonian is mathematically expressed as an operator acting on the Hilbert space of potential states of the system. In classical mechanics, the Hamiltonian is just a function mapping from the internal degrees to energy.

Intermolecular interactions   Molecules are aggregations of atoms that stick together via strong *intramolecular* interactions that can be broken apart only in chemical reactions (a single water molecule). *Intramolecular* interactions are relatively weaker forces between different molecules that determine the relative motion of molecules around each other (molecules in liquid water). Both intra- and intermolecular interactions are the end result of the single underlying Coulomb interaction between electrons and nuclei in molecules.

Many-body interactions   Effective models of systems composed of multiple interacting bodies (electrons, atoms, molecules) often describe collective property or behavior as resulting from a simple aggregate effect of the property or behavior of individual bodies and of pairs of bodies (pairwise interactions). In many cases, such an effective description captures a large part of the collective behavior. Collective behavior that cannot be expressed in terms of individual bodies and pairwise interactions is said to results from many-body interactions.

Materials design   One of two main branches of materials science that deals with discovery of novel materials with desired material properties. Sometimes also referred to as the inverse problem of material modeling. Computational materials design usually attempts to predict the atomic structure of the material with the desired properties, which can then be in principle prepared by synthetic experimental techniques.

Material modeling   One of two main branches of materials science that deals with prediction of properties of a given material. The general approach is to approximate the real materials with simplified model systems (with fewer degrees of freedom or simpler interactions), whose properties can be calculated using the laws of quantum and statistical mechanics.

Metastable state   The electrons in materials can be in different quantum states. Likewise the atoms in materials can be in different "states," which is a shorthand term for subsets of microstates that share some relevant physical feature. In both cases, if the system is allowed to interact with an environment that can "disturb" it, it can transition between the different states at any given moment with some probabilities. Some of the states are stable, which means that the probability of transitioning to any other state is low enough that it most likely does not happen on the relevant time scale. Unstable states are so short-lived, that the system never exhibits behavior that could be associated with any particular unstable state. Metastable states are between stable and unstable states in the sense that the system in a metastable state most likely transitions to other states during the relevant time scale, but it stays long enough in it that it exhibits behavior characteristic of that state.

Molecular dynamics   The atoms of a material constantly transition from one microstate to another, and the statistical distribution of the microstates determines many material properties. Molecular dynamics generates samples from this

distribution by evolving the positions of atoms along classical trajectories. The particular trajectories (sequences of microstates) are inconsequential, but the overall generated statistics can be used to calculate various thermodynamic properties.

Molecular geometry   Specification of the charges and positions of atomic nuclei in a molecule. Specifying the charges is equivalent to specifying the chemical identities of the atoms. The fixed nuclei are surrounded by the electronic cloud which determines the electronic properties of a molecule. The atoms of a molecule (the atomic nuclei) are in constant motion and a fixed molecular geometry corresponds to molecule frozen in time or a molecule at absolute zero temperature, when the atomic motion is greatly reduced.

Molecular symmetry   The geometry of many common molecules is symmetrical with respect to rotations, inversions, and reflections around a point. As a result of this symmetry, any observable function of the molecular geometry has to by invariant or equivariant with respect to these symmetry operations.

Observable   Physical laws operate with quantities. Some of these quantities can be in principle measured, and those are called physical observables, regardless of whether it is feasible to actually perform the measurement. Other quantities are only auxiliary intermediates used to formulate the physical laws, but cannot be measured directly. Examples of observables are distance, mass, square of a wave function, or an energy difference. Examples or non-observables are a wave function or absolute energy.

Periodic boundary conditions   One gram of a typical material contains on the order of 10^{21}–10^{23} atoms. To make their modeling tractable, a common approximation is to consider a sufficiently large box containing the atoms, which is then periodically repeated throughout the space. How large the box should be depends on the material and property in question. The errors caused by using a sufficiently large box are called finite size effects. The atomic structure of crystals is in fact periodic, so in the case of crystals periodic boundary conditions are not really an approximation.

Potential energy surface   The dependence of the energy of a molecule or a material on the positions of the atoms. The potential energy for the atoms is a result of the electronic motion and can be calculated by solving the Schrödinger equation of quantum mechanics. Each electronic state (ground state and excited states) has its own potential energy surface, which can cross. Such effects are important when studying the dynamics of excited states and electronic mechanisms of chemical reactions.

Quantum chemistry   Chemistry is the study of chemical reactions, and much of chemical knowledge preceded the discovery of quantum physics. Every since that discovery, quantum chemistry attempts to explain chemical properties of molecules from first principles using the tools of quantum mechanics. Quantum chemistry relies heavily on numerical calculations and the computational power of modern computers.

Schrödinger equation   The central eigenvalue equation of quantum mechanics that, given a specification of a system in the form of a Hamiltonian operator,

determines the possible quantum states in which the system can be found and the energies of those states. Depending on the basis in which the abstract operator equation is expressed, the Schrödinger equation can be either a differential equation or an algebraic matrix equation. Except for the simplest quantum-mechanical systems, the Schrödinger equation cannot be solved exactly, necessitating various approximations and numerical techniques.

References

1. J.G. Lee, *Computational Materials Science: An Introduction*, 2nd edn. (CRC Press, Boca Raton, 2017)
2. R. LeSar, *Introduction to Computational Materials Science: Fundamentals to Applications* (Cambridge University Press, Cambridge, 2013)
3. K. Ohno, K. Esfarjani, Y. Kawazoe, *Computational Materials Science: From Ab Initio to Monte Carlo Methods*, 2nd edn. (Springer, Berlin, 2018)
4. A.R. Leach, *Molecular Modelling: Principles and Applications*, 2nd edn. (Pearson Education, Edinburgh, 2001)
5. S. Yip, A. Wanda (eds.), *Handbook of Materials Modeling* (Springer, Berlin, 2020)
6. A.R. Oganov, G. Saleh, A.G. Kvashnin (eds.), *Computational Materials Discovery* (Royal Society of Chemistry, Cambridge, 2018)
7. L. Piela, *Ideas of Quantum Chemistry*, 2nd edn. (Elsevier, Amsterdam, 2014)
8. R.K. Pathria, P.D. Beale, *Statistical Mechanics*, 3rd edn. (Elsevier, Amsterdam, 2011)
9. L.D. Landau, E.M. Lifschitz, *Statistical Physics*, vol. 5, 3rd edn. (Pergamon Press, Oxford, 1980)
10. J.J. Sakurai, J. Napolitano, *Modern Quantum Mechanics*, 2nd edn. (Pearson Education, San Francisco, 2011)
11. R.P. Feynman, Phys. Rev. **56**(4), 340 (1939). https://doi.org/10.1103/PhysRev.56.340
12. D. Yoshioka, in *Statistical Physics* (Springer, Berlin, 2007), pp. 35–44