SCIENTIFIC REPORTS

OPEN

Accurate thermal conductivities from optimally short molecular dynamics simulations

Loris Ercole [1], Aris Marcolongo[2] & Stefano Baroni[1]

The evaluation of transport coefficients in extended systems, such as thermal conductivity or shear viscosity, is known to require impractically long simulations, thus calling for a paradigm shift that would allow to deploy state-of-the-art quantum simulation methods. We introduce a new method to compute these coefficients from optimally short molecular dynamics simulations, based on the Green-Kubo theory of linear response and the cepstral analysis of time series. Information from the *full* sample power spectrum of the relevant current for a *single* and relatively short trajectory is leveraged to evaluate and optimally reduce the noise affecting its zero-frequency value, whose expectation is proportional to the corresponding conductivity. Our method is unbiased and consistent, in that both the resulting bias and statistical error can be made arbitrarily small in the long-time limit. A simple data-analysis protocol is proposed and validated with the calculation of thermal conductivities in the paradigmatic cases of elemental and molecular fluids (liquid Ar and H_2O) and of crystalline and glassy solids (MgO and a-SiO_2). We find that simulation times of one to a few hundred picoseconds are sufficient in these systems to achieve an accuracy of the order of 10% on the estimated thermal conductivities.

Our microscopic understanding of heat and mass transport in extended systems is rooted in the Green-Kubo (GK) theory of linear response[1,2], as applied to the Navier-Stokes equations for the densities of the conserved extensive variables[3,4], which include energy, momentum, and the particle numbers for each molecular species. Transport coefficients are determined by the equilibrium fluctuations of the relevant currents and are in fact proportional to their autocorrelation times. Notwithstanding its beauty, rigor, and broad scope, the practical implementation of the GK theory is known to require very long molecular dynamics (MD) simulations, much longer in fact than the current autocorrelation times one is required to evaluate[5–9]. Even though a number of expedients have been devised to cope with this problem[7,9,10], none of them is fully satisfactory in that in some cases they fail to provide rigorous criteria to estimate the accuracy resulting from a given MD trajectory, and in all cases the length of the required simulations is unaffordable with accurate but expensive quantum simulation techniques[11].

It has long been believed that the GK theory of heat transport could not be combined with modern simulation methods based on electronic-structure theory because energy densities and currents are inherently ill-defined at the quantum-mechanical level[12]. This misconception has been recently challenged by a couple of papers showing that, while energy densities and currents are actually ill-defined classically no less than they are quantum-mechanically, the heat conductivity resulting from the GK formula is indeed well defined by virtue of a kind of *gauge invariance* stemming from energy extensivity and conservation[13,14]. This finding spurred a renewed interest in the quantum simulation of thermal conduction[11,15], and made it urgent to devise a data-analysis technique to make these simulations affordable, thus paving the way to the *ab initio* simulation of heat transport.

In this paper we propose a data-analysis protocol leading to an asymptotically unbiased and consistent estimate of transport coefficients (*i.e.* the bias and the statistical error can be made both arbitrarily small in the limit of long simulation times) and requiring shorter simulations than used so far. This protocol avoids any *ad hoc* fitting procedure and naturally provides an accurate estimate of the statistical error, thus lending itself to an easy implementation and automated use. While motivated by heat-transport applications, our approach naturally applies to *any* other transport properties that can be expressed, in a GK framework, in terms of time integrals of suitable autocorrelation functions, such as, *e.g.*, ionic conductivities, viscosities, and tracer diffusivity, to name

[1]SISSA – Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, 34136, Trieste, Italy. [2]Theory and Simulations of Materials (THEOS) and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland. Correspondence and requests for materials should be addressed to S.B. (email: baroni@sissa.it)

1

but a few. We test our procedure in the specific case of heat transport and we find that simulation times of one to a few hundred picoseconds are sufficient to achieve an accuracy of the order of 10% on the estimated thermal conductivities in a wide range of systems at typical physical conditions. In the Theory section we present the general theory of our approach; in the Applications and Benchmarks section we validate it with extensive benchmarks performed on the calculation of thermal conductivities in the paradigmatic cases of elemental and molecular fluids (liquid Ar and H_2O) and of crystalline and glassy solids (MgO and a-SiO_2). The Conclusions and Perspectives section finally contains our conclusions.

Theory

Heat transport in insulators is determined by the dynamics of atoms, the electrons following adiabatically in their ground state, and the GK theory of linear response[1,2] allows one to derive the thermal conductivity from an analysis of equilibrium, possibly *ab initio*, MD trajectories. According to this theory, the thermal conductivity of a macroscopic and isotropic system is given by:

$$\kappa = \frac{1}{Vk_BT^2} \int_0^\infty \langle \hat{J}(t)\hat{J}(0) \rangle \, dt,$$

(1)

where V and T are the system volume and temperature, $\hat{J}$ is one Cartesian component of the macroscopic heat flux[14,16], k_B is the Boltzmann constant, and the brackets $\langle \cdot \rangle$ indicate ensemble averages over initial conditions. Here and in the following a caret, as in "$\hat{J}$", denotes a realization of a random variable, whereas properties of the underlying stochastic process, such as time correlation functions, are indicated without carets. According to Eq. (1), the heat conductivity is proportional to the zero-frequency component of the power spectrum (aka the power spectral density) of the heat flux: $\kappa = \frac{1}{2Vk_BT^2}S(f=0)$,

$$S(f) = \int_{-\infty}^{\infty} e^{-i2\pi ft}G(t) \, dt,$$

(2)

where $G(t) = \langle \hat{J}(t)\hat{J}(0) \rangle$ is the equilibrium time-correlation function of the heat-flux process.

In practice, MD gives access to a discrete sample of the process, $\hat{J}_n = \hat{J}(n\varepsilon)$, where $0 \le n \le N-1$, ε is the sampling period of the current, and N the length of the time series, that we assume to be even. The Wiener-Khintchine theorem[17,18] states that, for large N, the power spectrum of the heat-flux process evaluated at $f_k = \frac{k}{N\varepsilon}$ is proportional to the expected value of the squared modulus of the discrete Fourier transform of the corresponding time series:

$$S(f_k) = \langle \hat{S}_k \rangle,$$

(3)

$$\hat{S}_k = \frac{\varepsilon}{N} |\hat{F}_k|^2,$$

(4)

$$\hat{F}_k = \sum_{n=0}^{N-1} e^{-2\pi i \frac{kn}{N}} \hat{J}_n,$$

(5)

for $0 \le k \le N-1$. The *sample spectrum* $\hat{S}_k$ is thus an estimator of the power spectrum of the process; in the statistics literature it is usually referred to as the *periodogram*. The reality of the $\hat{J}$'s implies that $\hat{F}_k = \hat{F}_{N-k}^*$ and $\hat{S}_k = \hat{S}_{N-k}$, so that periodograms are usually reported for $0 \le k \le \frac{N}{2}$ and their Fourier transforms evaluated as discrete cosine transforms.

The heat flux is an extensive quantity whose density autocorrelations are usually short-ranged. In the thermodynamic limit it can therefore be seen as the sum of (almost) independent identically distributed stochastic variables, so that, according to the central-limit theorem, its equilibrium distribution is Gaussian. A slight generalization of this argument allows us to conclude that the heat-flux process is Gaussian as well. The heat-flux time series is in fact a multivariate stochastic variable that, in the thermodynamic limit, results from the sum of (almost) independent variables, thus tending to a multivariate normal deviate. This implies that at equilibrium the real and imaginary parts of the $\hat{F}_k$'s defined in Eq. (5) are zero-mean normal deviates that, in the large-N limit, are uncorrelated among themselves and have variances proportional to the power spectrum evaluated at f_k. For $k=0$ or $k = \frac{N}{2}$, $\hat{F}_k$ is real and $\text{var}(\hat{F}_k) = \frac{N}{\varepsilon}S(f_k)$; for $k \notin \{0, \frac{N}{2}\}$, $\text{var}(\Re(\hat{F}_k)) = \text{var}(\Im(\hat{F}_k)) = \frac{N}{2\varepsilon}S(f_k)$. We conclude that in the large-N limit the sample spectrum of the heat-flux time series reads:

$$\hat{S}_k = S(f_k)\hat{\xi}_k,$$

(6)

where the $\hat{\xi}$'s are independent random variables distributed as a χ_1^2 variate for $k = 0$ or $k = \frac{N}{2}$ and as one half a χ_2^2 variate, otherwise. For the sake of simplicity, we make as though all the $\hat{\xi}_k$'s were identically distributed as one half a χ_2^2 variate for all values of k, thus making an error of order $\mathcal{O}(1/N)$, which vanishes in the long-time limit that is being assumed in most of the arguments presented in this paper. In many cases of practical interest, multiple time series are available to estimate the power spectrum of a process. For instance, in equilibrium MD, a same trajectory delivers one independent time series per Cartesian component of the heat flux, all of which are

obviously equivalent in isotropic systems. In these cases it is expedient to define a mean sample spectrum by averaging over ℓ different realizations:

$$^{\ell}\hat{S}_k = \frac{1}{\ell}\sum_{i=1}^{\ell}\hat{S}_k^i,$$

(7)

where the index "i" enumerates the realizations. For $k \notin \left\{0, \frac{N}{2}\right\}$ the ordinates of the mean sample spectrum are therefore distributed as in Eq. (6), with the $\hat{\xi}$'s being $\chi^2_{2\ell}$ variates, divided by the number of degrees of freedom:

$$^{\ell}\hat{\xi}_k \sim \frac{1}{2\ell}\chi^2_{2\ell}.$$

(8)

Eq. (6) brings both good and bad news. The good news is, the expectation of the periodogram of the time series is the power spectrum of the process, *i.e.* the former is an unbiased estimator of the latter. The bad news is, this estimator is not consistent, *i.e.* its variance does not vanish in the large-N limit. This is so because a longer time series increases the number of discrete frequencies at which the power spectrum is sampled, rather than its accuracy at any one of them. A consistent estimate of the zero-frequency value of the power spectrum (or the value at any other frequency, for that matter) can be obtained by segmenting a time series into several blocks of equal length and then averaging over the sample spectra computed for each of them. When the length of the trajectory grows large, so does the number of blocks, thus making the variance of the average arbitrarily small. In practice, the determination of the optimal block size is a unwieldy process that leads to an inefficient determination of the length of the trajectory needed to achieve a given overall accuracy. Here we adopt a different approach that allows us to obtain a consistent estimate of the zero-frequency value of the power spectrum from the statistical analysis of a *single* trajectory sample (*i.e.* no block analysis is needed) and such that the estimate of the trajectory length necessary to achieve a given accuracy is optimal.

Spectral density estimation from finite empirical time series is the subject of a vast literature in the statistical sciences, embracing both parametric and non-parametric methods[19]. In the following we propose a semi-parametric method to estimate the power spectrum of a stochastic process, based on a Fourier representation of the logarithm of its power spectrum (the "log-spectrum"). The advantage of dealing with the log-spectrum, instead of with the power spectrum itself, is twofold. First and foremost, the noise affecting the former is additive, instead of multiplicative, thus making it simple and expedient to apply linear filters: limiting the number of components of the Fourier representation of the log-spectrum acts as a low-pass filter that systematically reduces the power of the noise and yields a consistent estimator of the log-spectrum at any given frequency. Second, as a bonus, the logarithm is usually smoother than its argument. Therefore, the Fourier representation of the logarithm of the power spectrum is more parsimonious than that of the spectrum itself.

Let $^{\ell}\hat{L}_k = \log(^{\ell}\hat{S}_k)$ be the "*sample log-spectrum*" of our time series. By taking the logarithm of Eq. (6), $^{\ell}\hat{L}_k$ can be expressed as:

$$^{\ell}\hat{L}_k = \log(S(f_k)) + \log(^{\ell}\hat{\xi}_k) = \log(S(f_k)) + \lambda_\ell + ^{\ell}\hat{\lambda}_k,$$

(9)

where

$$\lambda_\ell = \int_0^\infty \log\left(\frac{\xi}{2\ell}\right)P_{\chi^2_{2\ell}}(\xi)\,d\xi = \psi(\ell) - \log(\ell)$$

(10)

is the expected value of the logarithm of the $^{\ell}\hat{\xi}$ stochastic variables defined in Eq. (8), $P_{\chi^2_{2\ell}}$ is the probability density of a $\chi^2_{2\ell}$ variate, $^{\ell}\hat{\lambda}_k = \log(^{\ell}\hat{\xi}_k) - \lambda_\ell$ are zero-mean identically distributed independent stochastic variables, and $\psi(z)$ and is the digamma function[20]. The variance of the $^{\ell}\hat{\lambda}$ variables is:

$$\sigma_\ell^2 = \int_0^\infty \log\left(\frac{\xi}{2\ell}\right)^2 P_{\chi^2_{2\ell}}(\xi)\,d\xi - \lambda_\ell^2 = \psi'(\ell),$$

(11)

where $\psi'(z)$ is the tri-gamma function[20].

Eq. (9) explicitly shows that the sample log-spectrum of a time series is equal to the logarithm of the power spectrum one wishes to evaluate (modulo a constant), *plus* a (non-Gaussian) white noise. In order to eliminate the high-frequency components of the noise, and thus reduce its total power, we define the "*cepstrum*" of the time series as the inverse Fourier transform of its sample log-spectrum:[21]

$$^{\ell}\hat{C}_n = \frac{1}{N}\sum_{k=0}^{N-1} {}^{\ell}\hat{L}_k e^{2\pi i\frac{kn}{N}},$$

(12)

and its coefficients as the *cepstral coefficients*. According to a generalized central-limit theorem for Fourier transforms of stationary time series[22], in the large-N limit, the (inverse) Fourier transform of the $^{\ell}\hat{\lambda}$'s, appearing in Eq. (9) and implicitly in Eq. (12), is a set of independent (almost) identically distributed zero-mean normal deviates. It follows that:

$$^{\ell}\hat{C}_n = \lambda_\ell\delta_{n0} + C_n + {}^{\ell}\hat{\mu}_n,$$

(13)

$$C_n = \frac{1}{N}\sum_{k=0}^{N-1}\log(S(f_k))e^{2\pi i \frac{kn}{N}}, \tag{14}$$

where ${}^{\ell}\hat{\mu}_n$ are independent zero-mean *normal* deviates with variances $\langle {}^{\ell}\hat{\mu}_n^2 \rangle = \frac{1}{N}\sigma_\ell^2$ for $n \notin \left\{0, \frac{N}{2}\right\}$ and $\langle {}^{\ell}\hat{\mu}_n^2 \rangle = \frac{2}{N}\sigma_\ell^2$ otherwise. This result can be easily checked explicitly by using the definition of the discrete Fourier transform; the non-trivial extra information provided by the central-limit theorem is the asymptotic independence and normality of the ${}^{\ell}\hat{\mu}$'s. Similarly to the sample power spectrum, the cepstral coefficients are real, periodic, and even: $\hat{C}_n = \hat{C}_{N-n}$. If the log-spectrum, $\log(S(f_k))$, is smooth enough, the number of non-negligible C_n coefficients in Eq. (14) is much smaller than N. Let us indicate with P^* the smallest integer such that $C_n \approx 0$ for $P^* \leq n \leq N - P^*$, and let us restrict the discrete Fourier transform of the cepstrum defined in Eq. (12) to $n < P^*$ or $n > N - P^*$. The $k=0$ component of this restricted Fourier transform reads:

$$ {}^{\ell}\hat{L}_0^* = {}^{\ell}\hat{C}_0 + 2\sum_{n=1}^{P^*-1}{}^{\ell}\hat{C}_n \tag{15}$$

$$ = \lambda_\ell + \log(S_0) + {}^{\ell}\hat{\mu}_0 + 2\sum_{n=1}^{P^*-1}{}^{\ell}\hat{\mu}_n. \tag{16}$$

We conclude that ${}^{\ell}\hat{L}_0^*$ is a normal deviate with expectation and variance

$$ \left\langle {}^{\ell}\hat{L}_0^* \right\rangle \equiv L_\ell^* = \log(S_0) + \lambda_\ell, \tag{17}$$

$$ \sigma_\ell^*(P^*, N)^2 = \sigma_\ell^2 \frac{4P^* - 2}{N}. \tag{18}$$

Eqs. (17) and (18) are the main results of our method: they show how the logarithm of the transport coefficient can be estimated from the (inverse) Fourier coefficients of the sample log-spectrum, Eq. (17), and how the resulting statistical error *only* depends on the number of Fourier coefficients retained and the total simulation length, Eq. (18).

The value of P^* is a property of the stochastic process underlying the time series, and is therefore independent of N: for any given value of P^* the variance σ_ℓ^{*2} tends to zero in the large-N limit, and ${}^{\ell}\hat{L}_0^* - \lambda_\ell$ is thus a consistent estimator of $\log(S_0)$. Notice that the absolute error on L_ℓ^* directly and nicely yields the relative error on S_0, which is proportional to the transport coefficient we are after. In general, all the cepstral coefficients are different from zero and assuming that many of them actually vanish introduces a bias. The optimal value of P^* is the one for which the bias, which is a decreasing function of it, is of the order of the statistical error, which instead increases with P^*. Its choice is the subject of *model selection* theory, another vast chapter in the statistical sciences[23]. Among the several tests that have been devised to perform this task, we choose the optimization of the Akaike's information criterion (AIC)[23,24], as described below.

Given a model depending on P parameters, $\theta = \{\theta_1, \theta_2, \cdots \theta_P\}$, the AIC is a sample statistic defined as

$$ \mathrm{AIC}(P) = -2\max_\theta \log \mathcal{L}(\theta, P) + 2P, \tag{19}$$

where $\mathcal{L}(\theta, P)$ is the likelihood of the parameters. The optimal number of parameters is then determined as the argument of the AIC minimum:

$$ P_A^* \equiv \underset{P}{\mathrm{argmin}}\ \mathrm{AIC}(P). \tag{20}$$

In the present case the parameters of the model are the P coefficients $C = \{C_0, C_1, \cdots C_{P-1}\}$ as defined in Eq. (14), and the log-likelihood reads, up to additive terms independent of P and C:

$$ 2\log\mathcal{L}(C, P) = -\frac{N}{2\sigma_\ell^2}(C_0 + \lambda_\ell - \hat{C}_0)^2 - \frac{N}{\sigma_\ell^2}\sum_{n=1}^{P-1}(C_n - \hat{C}_n)^2 - \frac{N}{\sigma_\ell^2}\sum_{n=P}^{\frac{N}{2}}\hat{C}_n^2. \tag{21}$$

Evidently, the above expression is maximized, for given P, by $C_n = \hat{C}_n - \delta_{n0}\lambda_\ell$ for $n = 0, 1, \cdots P-1$, and the corresponding value of the maximum is: $2\max_C \log\mathcal{L}(C, P) = -\frac{N}{\sigma_\ell^2}\sum_{n=P}^{\frac{N}{2}}\hat{C}_n^2$. We conclude that the value of the AIC is:

$$ \mathrm{AIC}(P) = \frac{N}{\sigma_\ell^2}\sum_{n=P}^{\frac{N}{2}}\hat{C}_n^2 + 2P. \tag{22}$$

The value of P that minimizes this expression is the optimal number of parameters in the Akaike sense, P_A^*, as defined in Eq. (20).

The maximum frequency available for spectral/cepstral analysis is the Nyqvist frequency[25], determined by the sampling period ε as $f_{\mathrm{Ny}} = \frac{1}{2\varepsilon}$. Transport coefficients only depend on the low-frequency behavior of the spectrum, which is independent of ε, as long as the latter is small enough as to avoid aliasing effects. For this reason it may prove convenient to eliminate the high-frequency portion of the spectrum ($f > f^*$) by applying a low-pass filter to the time series (*e.g.* a moving average[26]) and then resample the latter with a sampling period $\varepsilon^* = \frac{1}{2f^*}$, thus resulting in a time series of $N^* = N\frac{f^*}{f_{\mathrm{Ny}}}$ time steps. The optimal number of cepstral coefficients resulting from Eqs. (20) and (22), as well as the error in the estimate of the transport coefficients resulting from Eq. (18), depends in general on the choice of the cutoff frequency, f^*. The smaller f^*, the smaller will presumably be the number of cepstral coefficients necessary to describe the log-spectrum to any given accuracy over such a shorter frequency range. However, the shorter length of the filtered time series, N^*, results in an increased variance of the estimator $^\ell\hat{L}_0^*$ defined in Eq. (15), according to Eq. (18). Numerical experiments performed on the MD data reported in the Applications and Benchmarks section show that both the estimated value of $^\ell\hat{L}_0^*$ and its variance are actually fairly insensitive to the value chosen for the cutoff frequency, f^*, provided that the power spectrum of the re-sampled time series faithfully features the first band of the original spectrum (*i.e.* the first prominent feature) and that this band is not too peaked at the origin.

Data analysis based on the theory presented above is straightforward. Given a current time series from a MD trajectory, the sample spectrum $^\ell\hat{S}_k$ is first determined from Eqs. (4), (5), and (7) and smoothed out using *e.g.* a moving average[26] performed over a narrow frequency window, so as to reduce the statistical noise to a level where the shape of the power spectrum of the underlying process can be appreciated. A cutoff frequency, f^*, is then determined so as to encompass the first band of the smoothed power spectrum. The cepstrum $^\ell\hat{C}_n$ is computed by Fourier analyzing the logarithm of the sample spectrum up the cutoff frequency thus estimated, as in Eq. (12), and the optimal number of cepstral coefficients is determined using Eqs. (20) and (22). The logarithm of the zero-frequency value of the power spectrum is finally estimated from Eqs. (15) and (17) as $^\ell\hat{L}_0^* - \lambda_\ell$, and its variance from Eq. (18).

Mind the difference between the moving average performed in the frequency domain to smooth out the power spectrum and that performed in the time domain, as suggested before, and acting as a low-pass filter. Spectral smoothing using a moving average in the frequency domain is common practice in the analysis of time series, and it actually provides a consistent estimator of the power spectrum. In fact, the number of frequencies falling within a window of given width increases linearly with the length of the series, so that the variance of the average decreases as the inverse of the product of the length of the series times the width of the window. The resulting spectral estimate is however biased by the variation of the signal within the window, thus strongly reducing the width of the windows that can be afforded. Moreover, both the bias and the statistical error are difficult to estimate due to the multiplicative nature of the noise; therefore neither the plain periodogram nor a running average thereof are adequate for a quantitative estimate of the zero-frequency value of the power spectrum, which is proportional to the transport coefficient we are after.

Applications and Benchmarks

In order to benchmark the methodology described above, we have applied it to the calculation of the thermal conductivity in four systems representative of different classes of materials, spanning from elemental and molecular fluids (Ar and H_2O) to crystalline and glassy solids (MgO and a-SiO$_2$). Classical MD simulations were run using the LAMMPS package[27] with the setup described below. For liquid Ar we used a Lennard Jones potential as described in Ref.[28] at a temperature $T \approx 220\,K$ and density $\rho = 1.55\,g/cm^3$, in a cubic supercell containing 864 atoms, with a time step $\Delta t = 4\,fs$. For liquid H_2O we used a flexible model as in Ref.[29] at a temperature $T \approx 300\,K$ and density $\rho = 1.0\,g/cm^3$, in a cubic supercell containing 180 molecules, with a time step $\Delta t = 0.5\,fs$. For crystalline (FCC) MgO we used a Buckingham-plus-Coulomb potential as in Ref.[30] at a temperature $T \approx 1000\,K$ and density $\rho = 3.61\,g/cm^3$, in a $4\times4\times4$ simple cubic conventional supercell with 512 atoms, and a time step $\Delta t = 0.3\,fs$. For a-SiO$_2$ we used the potential proposed by van Beest, Kramer, and van Santen[31] as implemented by Shcheblanov *et al.*[32], at a temperature $T \approx 1000\,K$ and density $\rho = 2.29\,g/cm^3$, in a supercell containing 216 atoms with a time step $\Delta t = 1\,fs$. The glass model was obtained from a quench from the melt ($T \approx 6500\,K$) at a constant quenching rate of $5.5 \times 10^{12}\,K/s$. Each system was equilibrated in the NVT ensemble at the target temperature for several hundred picoseconds; data were then collected in the NVE ensemble for trajectories whose length was chosen so as to represent realistic simulation runs that could be afforded using *ab initio* MD ($\mathcal{T} = 100\,ps$ for Ar, H_2O, and a-SiO$_2$, and $\mathcal{T} = 500\,ps$ for MgO). In order to compare our estimates of the transport coefficients and their statistical errors with reliable and statistically significant reference data, in all cases we ran much longer ($\approx 50\,ns$) simulations. This allowed us to compare our predicted conductivities with accurate values estimated from the direct integration of the time auto-correlation function in Eq. (1), as obtained from a block average[33] performed over the long trajectory. In addition, we could collect abundant statistics of our estimator for the transport coefficients, Eq. (15), and validate its normal distribution specified by Eqs. (17) and (18).

In Fig. 1 we report the sample periodograms of the heat fluxes computed for the four systems considered in this paper and averaged over the three Cartesian components, as described above. The solid lines in color indicate a further moving average[26] computed over a narrow frequency window, as explained at the end of the Theory section. The values of the cutoff frequencies used for cepstral analysis, f^*, are chosen so as to encompass the first prominent feature of the (smoothed) power spectrum. For instance, in H_2O and a-SiO$_2$ we choose $f^* \approx 29\,THz$ and $f^* \approx 28\,THz$, respectively. In MgO we assume $f^* \approx 0.6\,THz$, just at the upper edge of the first narrow peak,
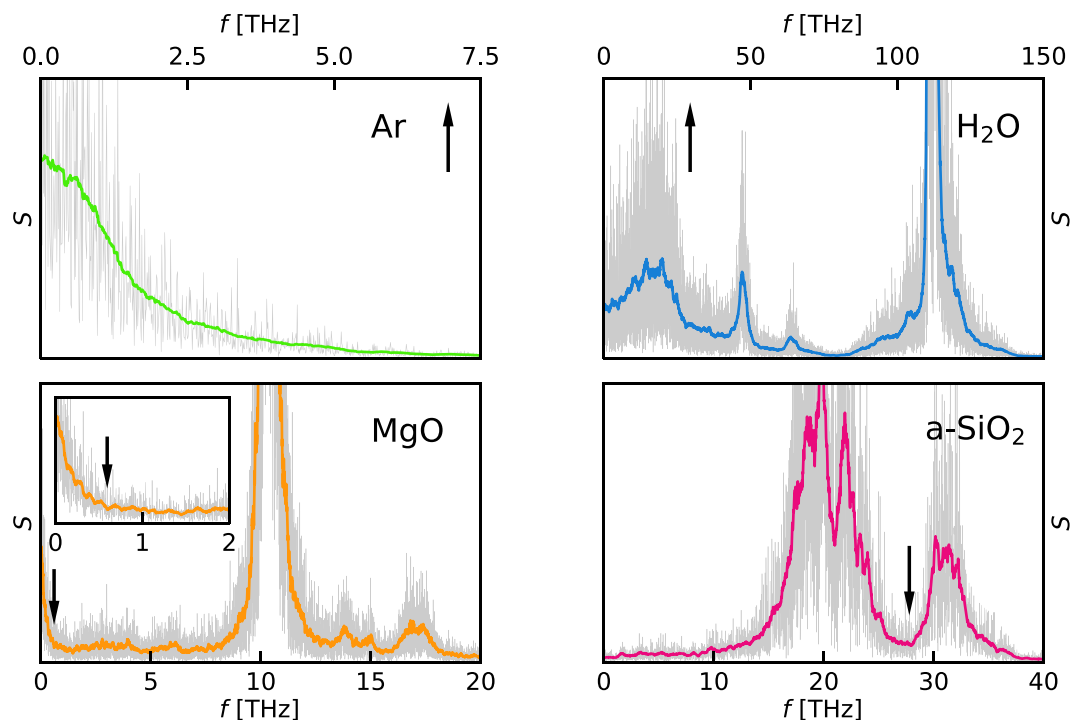
Figure 1. Sample power spectra of the heat flux computed from MD trajectories for Ar, H_2O, a-SiO_2 (100 ps) and MgO (500 ps) (gray line, see text). The solid lines in color correspond to a moving average performed over a narrow frequency window. The vertical arrows indicate the cutoff frequencies, f^*, used for the subsequent cepstral analysis (see text). The inset in the MgO panel is a magnification of the low-frequency region of the spectrum.

whereas in Ar there is just one band, corresponding to a purely diffusive behavior of a simple fluid, and we assume $f^* \approx 7$ THz, where the spectrum has exhausted most of the available power, but its value is not yet too small (see Fig. 1). The corresponding average numbers of cepstral coefficients given by the optimization of the AIC are: $P_A^* = 5$ (Ar), 7 (H_2O), 4 (MgO), and 31 (a-SiO_2). Later we will will display the dependence of the number of optimal cepstral coefficients and of the resulting estimate of the thermal conductivity on the choice of f^*, and show that this choice is not critical.

In order to validate our data-analysis protocol, we first computed the heat conductivities from a direct integration of the current autocorrelation function, Eq. (1), combined with standard block analysis over the 50 ns long trajectory, that will be taken as a reference, obtaining: $\kappa_{ref} = 0.1965 \pm 0.0015, 0.970 \pm 0.009, 19.2 \pm 0.4$, and 2.115 ± 0.025 W/mK, for Ar, H_2O, MgO, and a-SiO_2, respectively. Although our simulations were meant for benchmarking purposes only, and no particular care was paid to exactly match the simulation conditions of previous work, these data are in fair agreement with the foregoing theoretical results: ≈ 0.19 W/mK (Ar)[28], ≈ 0.85 W/mK (H_2O)[34], ≈ 12 W/mK (MgO)[30], and ≈ 2.1 W/mK (a-SiO_2)[35].

In Fig. 2 we display the distributions of the values of $\log(\kappa/\kappa_{ref})$ estimated by applying our protocol to multiple MD segments of 100 ps (for Ar, H_2O, a-SiO_2) and 500 ps (for MgO), extracted from the 50 ns long trajectory. The optimal numbers of cepstral coefficients, P^*, have been redetermined for each segment independently, while the values of the cutoff frequency, f^*, which only depends on the qualitative features of the spectrum, have been determined once for all for one of them. The distribution of the resulting number of cepstral coefficients is reported in Fig. 3. The observed distributions of $\log(\kappa)$ successfully pass the Shapiro-Wilk normality test[36] (*i.e.* they do not fail it) and the observed sample standard deviations closely match the theoretical values estimated from Eq. (18) (sample/theory): 0.104/0.079 (Ar), 0.053/0.045 (H_2O), 0.17/0.11 (MgO), and 0.114/0.095 (a-SiO_2). Remember that the error on $\log(\kappa)$ is the relative error on κ: the corresponding absolute errors achievable with a short trajectory of 100 (Ar, H_2O, and a-SiO_2) or 500 (MgO) ps, not to be confused with the long trajectory used to establish the reference data above, are therefore: $\sigma_\kappa \approx 0.015$ (Ar), 0.045 (H_2O), 2 (MgO), and 0.2 (a-SiO_2) W/mK. This indicates that a *single* and short sample trajectory, such as one that is affordable with *ab initio* MD, is sufficient to achieve and accurately estimate a very decent relative error on the computed transport coefficient. In an attempt to evaluate the thermal conductivity from the direct computation of the GK integral, Eq. (1), and standard block analysis using similarly short MD trajectories, our best estimate of the resulting statistical error was 2–3 times larger than using our protocol (meaning 5–10× longer trajectories to achieve a comparable accuracy) in all cases but liquid Ar, where only a marginal improvement is achieved using our methodology. Much more than this, the standard analysis of MD data depends on a number of hidden parameters, such as the upper limit of the GK integral, Eq. (1), or the width of the blocks for error analysis, that are hard to determine and keep under control. Our method, instead, only depends on a single parameter, the number of cepstral coefficients, whose optimal
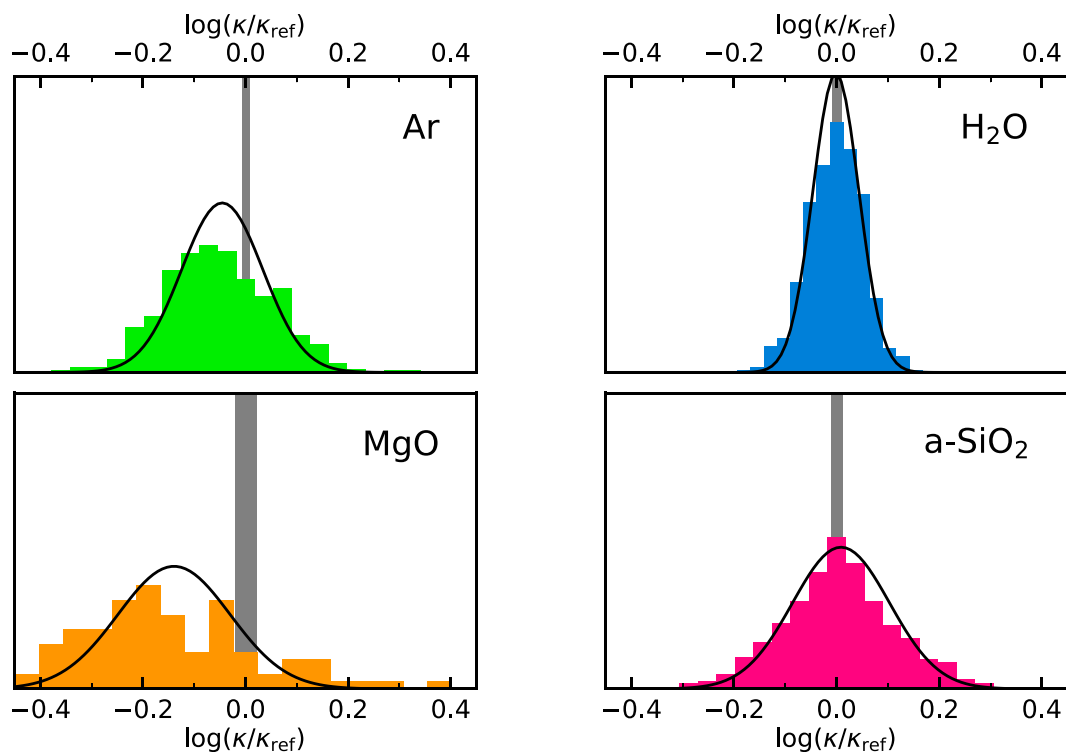
Figure 2. Distributions of the logarithm of the thermal conductivities, $\log(\kappa)$, estimated over multiple MD segments (100 ps for Ar, H_2O, and a-SiO_2, and 500 ps for MgO) extracted from a 50 ns long trajectory. The reported data are referred to κ_{ref}, which is the value obtained from the direct integration of the current autocorrelation function in Eq. (1), combined with standard block analysis over the 50 ns trajectory, and represented by the vertical gray bands. The Gaussian curves represent the distributions predicted by the theory, centered at the sample mean. Remember that the absolute error on $\log(\kappa)$ is the relative error on κ.
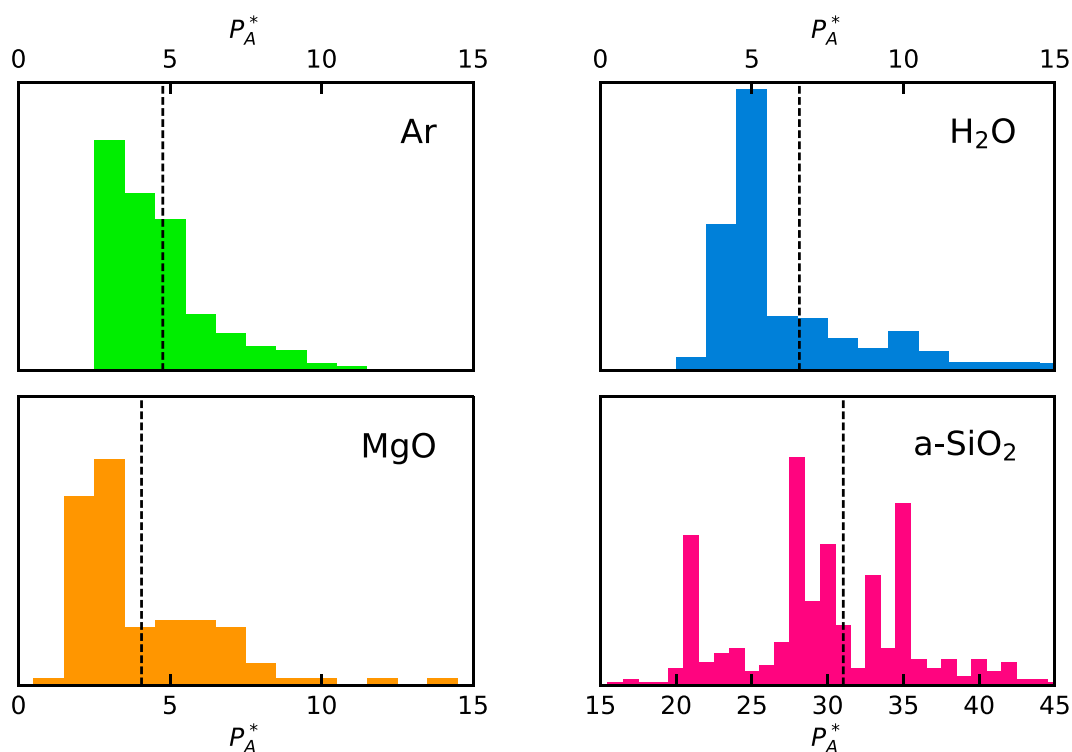
Figure 3. Distribution of the optimal numbers of cepstral coefficients, P_A^*, as determined by optimizing the Akaike's information criterion, Eqs. (20) and (22), for each segment of the 50 ns long MD trajectory, as described in the text. The vertical dashed lines indicate the average value of P_A^*.
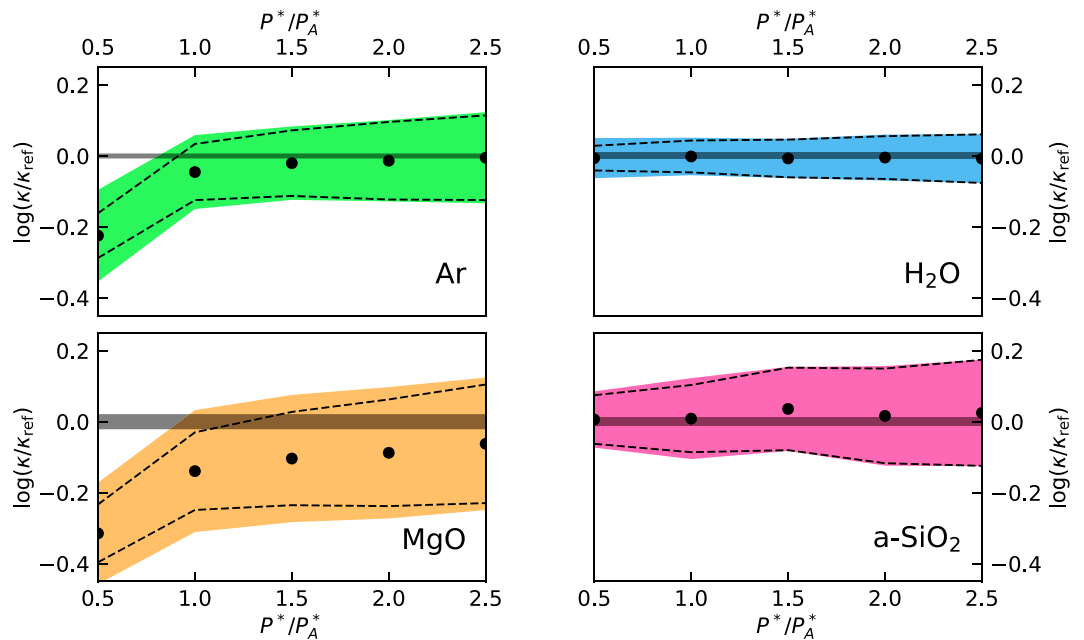
Figure 4. Dependence of $\log(\kappa)$, as estimated from Eq. (15), on the number of cepstral coefficients P^*. P_A^* is the optimal number of coefficients estimated from the AIC using Eqs. (20) and (22). The black dots represent the mean values of $\log(\kappa)$ computed over multiple MD segments (100 ps for Ar, H_2O, and a-SiO_2, and 500 ps for MgO) extracted from a 50 ns long trajectory; the colored bands and dashed lines represent one standard deviation as estimated from the empirical statistics and from Eq. (18), respectively. The reported data are referred to κ_{ref}, which is the value of thermal conductivity obtained from direct integration of the current autocorrelation function in Eq. (1), combined with standard block analysis over the 50 ns trajectory, and represented by the horizontal gray bands. Remember that the absolute error on $\log(\kappa)$ is the relative error on κ.

value can be easily determined from the Akaike's information criterion, or other more sophisticated model-selection methods[23,37], as appropriate.

In order to estimate the bias introduced by limiting the number of cepstral coefficients, we examined the sample mean computed over the distributions displayed in Fig. 2, $\langle\kappa\rangle$, obtaining: $\langle\kappa\rangle = 0.1878 \pm 0.0007$, $0.969 \pm 0.002, 16.7 \pm 0.2$, and 2.131 ± 0.009 W/mK, for Ar, H_2O, MgO, and a-SiO_2, respectively. Comparing these data with the reference data obtained from the direct evaluation of the GK integral, we see that the bias is negligible for H_2O and a-SiO_2, very small for Ar, and small but not negligible for MgO. In Fig. 4 we display the dependence of $\log(\kappa/\kappa_{ref})$ on the number of cepstral coefficients, P^*, as estimated from Eq. (15). We observe that when the number of cepstral coefficients, P^*, is larger than the optimal value determined from the AIC, P_A^*, the estimated value of κ seems not to depend on P^* for all systems but MgO, for which a slight bias seems to persist, and to a much lesser extent for Ar. Also, Eq. (18) seems to slightly underestimate the sample variance for small P^* in these cases. In the case of MgO this behavior is likely due to the difficulty of the AIC to cope with the sharp low-frequency peak in the power spectrum, due to the highly harmonic character and slow decay of the vibrational heat carriers in periodic crystals[11], thus requiring longer simulation times. In the case of Ar the very small bias observed for $P^* = P_A^*$ may be due to the difficulty of choosing a suitable cutoff frequency when only a single diffusive band is present in the spectrum, and to the divergence of the log-spectrum at high frequency. In all cases, use of the Aikake's information criterion results in a bias that is smaller than the statistical error estimated from an individual short sample trajectory and that can be systematically removed by increasing the value of P^*, at the price of increasing the statistical error, if and when needed.

Finally, in Fig. 5 we report the dependence of the optimal number of cepstral coefficients, P_A^*, as a function of the cutoff frequency, f^*, along with the dependence of the resulting estimate of $\log(\kappa/\kappa_{ref})$. P_A^* increases with f^*. Notwithstanding, the estimated value of the heat conductivity, as well as its variance, is fairly insensitive on the precise value of f^* as long as the latter is large enough as to encompass the lowermost prominent feature of the spectrum. Some more comments are in order for MgO. In this case the high thermal conductivity, due to the strong harmonic character of slowly decaying phonon modes, manifests in the form of a narrow peak centered at $f = 0$, followed by a broad plateau that carries little spectral weight. This feature determines a more pronounced increase in the number of significant cepstral coefficients as f^* increases and a corresponding increase of the bias when keeping P^* at the value given by the AIC. The AIC is a less reliable indicator of the number of cepstral coefficients necessary to keep the bias low, in this case. By increasing this number by a factor of two or more, the bias decreases, as indicated by the results reported in lighter colors in Fig. 5, and eventually vanishes, as shown in Fig. 4.
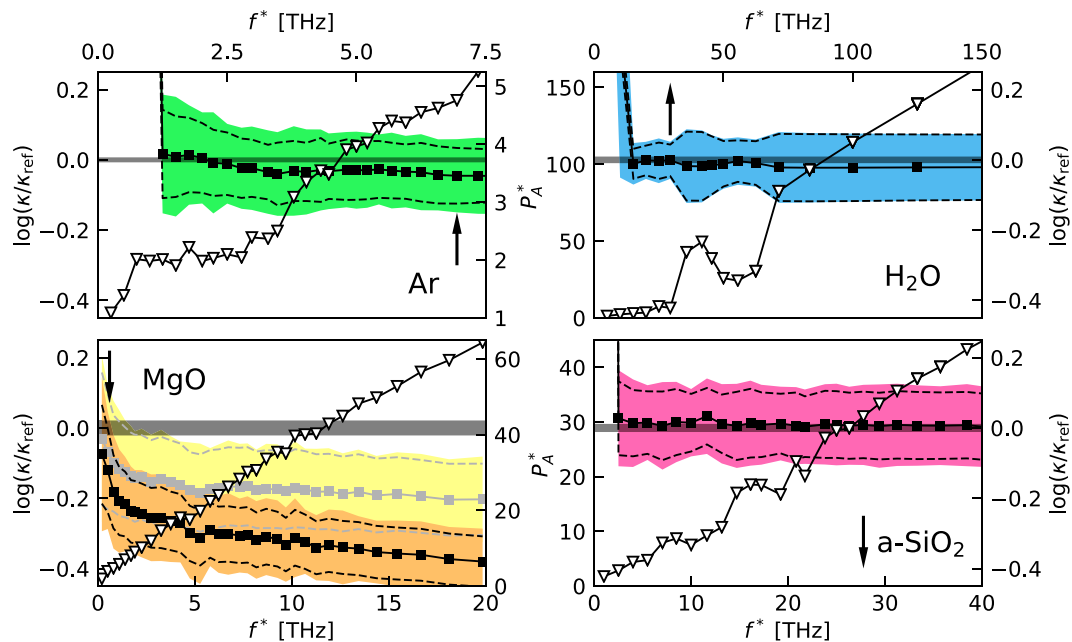
Figure 5. Triangles: average optimal number of cepstral coefficients, P_A^*, as determined by the AIC, Eqs. (20) and (22), as a function of the cutoff frequency used for cepstral analysis, f^* (see discussion just after Eq. (22)). Squares: $\log(\kappa)$ resulting from a given choice of f^* and of the corresponding value of P_A^*. All the values are averages performed over multiple 100 ps long segments (500 ps for MgO) extracted from a 50 ns long MD trajectory, as discussed in the text. The colored bands indicate the sample standard deviation and the dashed lines that resulting from our theoretical analysis (see Eq. (18)). The vertical arrows indicate the cutoff frequencies, f^*, used for the cepstral analysis in this paper (see Fig. 1 and text). In the case of MgO, the data indicated with lighter colors are obtained using a number of cepstral coefficients twice as large as that provided by the AIC, $P^* = 2P_A^*$. The data are referred to κ_{ref}, which is the value of thermal conductivity obtained from direct integration of the current autocorrelation function in Eq. (1), combined with standard block analysis over the 50 ns trajectory, and represented by the horizontal gray bands. Remember that the absolute error on $\log(\kappa)$ is the relative error on κ.

Conclusions and Perspectives

We believe that the methodology introduced in this paper will finally open the way to the quantum simulation of heat transport, which has been considered out of its scope until very recently, particularly for strongly anharmonic and/or disordered systems. Its implementation is straightforward and its use robust, as the only parameter to be determined is the optimal number of cepstral coefficients, using *e.g.* the Akaike's information criterion. Our methodology performs best when a low conductivity results from large cancellations in the integral of a highly oscillatory time auto-correlation function, such as in molecular liquids or amorphous solids, where simulation times of the order of 100 ps seem sufficient to obtain accuracies of the order of 10% in the estimated thermal conductivities. The performance is less spectacular in periodic crystals, where slowly-decaying strongly-harmonic phonon modes require longer simulation times and the ensuing sharp peak in the low-frequency region of the power spectrum requires a larger number of cepstral coefficients than predicted by the optimization of the AIC. Even so, simulation times of the order of a few hundred picoseconds seem sufficient to achieve a comparable accuracy. In the latter case, it is possible that a combination of the methodology introduced here with specialized techniques based on normal-mode analysis, such as that presented in Ref.[11], will result in further improvements. Replacing the optimization of the AIC with more sophisticated and possibly more efficient approaches, such as *e.g.* weighted multi-model inference techniques[23,37], may also assist in this and other difficult cases. Finally, we expect that our methodology will impact on the simulation of any transport phenomena to which the Green-Kubo theory applies, such as ionic conduction, viscosity, and many others.

References

1. Green, M. S. Markoff Random Processes and the Statistical Mechanics of Time-Dependent Phenomena. II. Irreversible Processes in Fluids. *J. Chem. Phys.* **22**, 398–413, https://doi.org/10.1063/1.1740082 (1954).
2. Kubo, R. Statistical-Mechanical Theory of Irreversible Processes. I. General Theory and Simple Applications to Magnetic and Conduction Problems. *J. Phys. Soc. Jpn.* **12**, 570–586, https://doi.org/10.1143/JPSJ.12.570 (1957).
3. Kadanoff, L. P. & Martin, P. C. Hydrodynamic equations and correlation functions. *Ann. Phys.* **24**, 419–469, https://doi.org/10.1016/0003-4916(63)90078-2 (1963).
4. Forster, D. *Hydrodynamic fluctuations, broken symmetry, and correlation functions* (Reading, Mass., W. A. Benjamin, Inc., 1975).
5. Schelling, P. K., Phillpot, S. R. & Keblinski, P. Comparison of atomic-level simulation methods for computing thermal conductivity. *Phys. Rev. B* **65**, 1–12, https://doi.org/10.1103/PhysRevB.65.144306 (2002).
6. Nevins, D. & Spera, F. J. Accurate computation of shear viscosity from equilibrium molecular dynamics simulations. *Mol. Simul.* **33**, 1261–1266, https://doi.org/10.1080/08927020701675622 (2007).

7. Jones, R. E. & Mandadapu, K. K. Adaptive Green-Kubo estimates of transport coefficients from molecular dynamics based on robust error analysis. *J. Chem. Phys.* **136**, 154102, https://doi.org/10.1063/1.3700344 (2012).

8. Zhang, Y., Otani, A. & Maginn, E. J. Reliable Viscosity Calculation from Equilibrium Molecular Dynamics Simulations: A Time Decomposition Method. *J. Chem. Theory Comput.* **11**, 3537–3546, https://doi.org/10.1021/acs.jctc.5b00351 (2015).

9. Oliveira, L. D. S. & Greaney, P. A. Method to manage integration error in the Green-Kubo method. *Phys. Rev. E* **95**, 023308, https://doi.org/10.1103/PhysRevE.95.023308 (2017).

10. Chen, J., Zhang, G. & Li, B. How to improve the accuracy of equilibrium molecular dynamics for computation of thermal conductivity? *Phys. Lett. A* **374**, 2392–2396, https://doi.org/10.1016/j.physleta.2010.03.067 (2010).

11. Carbogno, C., Ramprasad, R. & Scheffler, M. Ab initio Green-Kubo Approach for the Thermal Conductivity of Solids. *Phys. Rev. Lett.* **118**, 175901, https://doi.org/10.1103/PhysRevLett.118.175901 (2017).

12. Stackhouse, S., Stixrude, L. & Karki, B. B. Thermal Conductivity of Periclase (MgO) from First Principles. *Phys. Rev. Lett.* **104**, 208501, https://doi.org/10.1103/PhysRevLett.104.208501 (2010).

13. Marcolongo, A., Umari, P. & Baroni, S. Microscopic theory and ab initio simulation of atomic heat transport. *Nat. Phys.* **12**, 80–84, https://doi.org/10.1038/nphys3509 (2016).

14. Ercole, L., Marcolongo, A., Umari, P. & Baroni, S. Gauge Invariance of Thermal Transport Coefficients. *J. Low Temp. Phys.* **185**, 79–86, https://doi.org/10.1007/s10909-016-1617-6 (2016).

15. Kang, J. & Wang, L. First-principles green-kubo method for thermal conductivity calculations. *Phys. Rev. B* **96**, 020302, https://doi.org/10.1103/PhysRevB.96.020302 (2017).

16. Hansen, J.-P. & McDonald, I. R. *Theory of Simple Liquids with Applications to Soft Matter*, 4th edn (Academic Press, Oxford, 2013).

17. Wiener, N. Generalized harmonic analysis. *Acta Math.* **55**, 117–258, https://doi.org/10.1007/BF02546511 (1930).

18. Khintchine, A. Korrelationstheorie der stationären stochastischen Prozesse. *Math. Ann.* **109**, 604–615, https://doi.org/10.1007/BF01449156 (1934).

19. Stoica, P. S. & Moses, R. *Spectral Analysis of Signals* (Prentice Hall, Upper Saddle River, 2005).

20. Weisstein, E. W. *Polygamma Function*, from MathWorld – a Wolfram Web Resource, http://mathworld.wolfram.com/PolygammaFunction.html.

21. Childers, D. G., Skinner, D. P. & Kemerait, R. C. The Cepstrum: A Guide to Processing. *Proc. IEEE* **65**, 1428–1443, https://doi.org/10.1109/PROC.1977.10747 (1977).

22. Peligrad, M. & Wu, W. B. Central limit theorem for fourier transforms of stationary processes. *Ann. Prob.* **38**, 2009–2022, https://doi.org/10.1214/10-AOP530 (2010).

23. Claeskens, G. & Hjort, N. L. *Model Selection and Model Averaging* (Cambridge University Press, 2008).

24. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **19**, 716–723, https://doi.org/10.1109/TAC.1974.1100705 (1974).

25. Oppenheim, A. V., Schafer, R. W. & Buck, J. R. *Discrete-Time Signal Processing* (*2nd Edition*) (*Prentice-Hall Signal Processing Series*) (Prentice Hall, 1999).

26. Weisstein, E. W. *Moving Average*, from MathWorld – a Wolfram Web Resource, http://mathworld.wolfram.com/MovingAverage.html.

27. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.* **117**, 1–19, https://doi.org/10.1006/jcph.1995.1039 (1995).

28. Tretiakov, K. V. & Scandolo, S. Thermal conductivity of solid argon from molecular dynamics simulations. *J. Chem. Phys.* **120**, 3765–3769, https://doi.org/10.1063/1.1642611 (2004).

29. Alejandre, J., Chapela, G. A., Bresme, F. & Hansen, J. P. The short range anion-H interaction is the driving force for crystal formation of ions in water. *J. Chem. Phys.* **130**, https://doi.org/10.1063/1.3124184 (2009).

30. Shukla, P., Watanabe, T., Nino, J. C., Tulenko, J. S. & Phillpot, S. R. Thermal transport properties of MgO and Nd2Zr2O7 pyrochlore by molecular dynamics simulation. *J. Nucl. Mat.* **380**, 1–7, https://doi.org/10.1016/j.jnucmat.2008.06.043 (2008).

31. van Beest, B. W. H., Kramer, G. J. & van Santen, R. A. Force fields for silicas and aluminophosphates based on ab initio calculations. *Phys. Rev. Lett.* **64**, 1955, https://doi.org/10.1103/PhysRevLett.64.1955 (1990).

32. Shcheblanov, N. S., Mantisi, B., Umari, P. & Tanguy, A. Detailed analysis of plastic shear in the Raman spectra of SiO2 glass. *J. Non-Cryst. Solids* **428**, 6–19, https://doi.org/10.1016/j.jnoncrysol.2015.07.035 (2015).

33. Frenkel, D. & Smit, B. *Understanding Molecular Simulation*: *From Algorithms to Applications* (*Computational Science Series*, *Vol 1*) (Academic Press, 2001).

34. Römer, F., Lervik, A. & Bresme, F. Nonequilibrium molecular dynamics simulations of the thermal conductivity of water: A systematic investigation of the SPC/E and TIP4P/2005 models. *J. Chem. Phys.* **137**, 074503, https://doi.org/10.1063/1.4739855 (2012).

35. Larkin, J. M. & McGaughey, A. J. H. Thermal conductivity accumulation in amorphous silica and amorphous silicon. *Phys. Rev. B* **89**, 144303, https://doi.org/10.1103/PhysRevB.89.144303 (2014).

36. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biom.* **52**, 591, https://doi.org/10.2307/2333709 (1965).

37. Burnham, K. P. & Anderson, R. P. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Soc. Meth. Res.* **33**, 261–304, https://doi.org/10.1177/0049124104268644 (2004).

Acknowledgements

Author Contributions

L.E. contributed to the development of the theory, ran most of the simulations, wrote the data-analysis software, performed the data analysis, and contributed to write the manuscript. A.M. contributed to the development of the theory, ran some of the simulations, and contributed to write the manuscript. S.B. developed the theory, supervised the project, and wrote the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.