LETTER

Characterising particulate random media from near-surface backscattering: A machine learning approach to predict particle size and concentration

To cite this article: Artur L. Gower *et al* 2018 *EPL* **122** 54001

View the article online for updates and enhancements.

Related content

Characterising particulate random media from near-surface backscattering: A machine learning approach to predict particle size and concentration

Artur L. Gower[1], Robert M. Gower[2], Jonathan Deakin[1], William J. Parnell[1] and I. David Abrahams[3]

[1] *School of Mathematics, University of Manchester - Oxford Road, Manchester, M13 9PL, UK*
[2] *LTCI, Télécom Paristech, Université Paris-Saclay - 75013, Paris, France*
[3] *Isaac Newton Institute for Mathematical Sciences - 20 Clarkson Road, Cambridge, CB3 0EH, UK*

Abstract – To what extent can particulate random media be characterised using direct wave backscattering from a single receiver/source? Here, in a two-dimensional setting, we show using a machine learning approach that both the particle radius and concentration can be accurately measured when the boundary condition on the particles is of Dirichlet type. Although the methods we introduce could be applied to any particle type. In general backscattering is challenging to interpret for a wide range of particle concentrations, because multiple scattering cannot be ignored, except in the very dilute range. Across the concentration range from 1% to 20% we find that the mean backscattered wave field is sufficient to accurately determine the concentration of particles. However, to accurately determine the particle radius, the second moment, or average intensity, of the backscattering is necessary. We are also able to determine what is the ideal frequency range to measure a broad range of particles sizes. To get rigorous results with supervised machine learning requires a large, highly precise, dataset of backscattered waves from an infinite half-space filled with particles. We are able to create this dataset by introducing a numerical approach which accurately approximates the backscattering from an infinite half-space.

Under close inspection, many materials are composed of small randomly distributed particles or inclusions. So it is no surprise that the need to measure particle properties, such as their average size and concentration, spans many physical disciplines. For quick non-invasive measurements, waves, either mechanical, electromagnetic or quantum, are the preferred choice. However, measuring a broad range of particle concentrations and sizes is still an open challenge. For high concentrations the wave undergoes multiple scattering, which requires specialised methods to compute and interpret. And further, measuring a wide range of particle sizes means a wide range of frequencies needs to be considered.

The type of wave used depends on the type of particle: acoustic waves are used to measure liquid emulsions [1], sediment on the ocean floor [2] and polycrystalline materials [3]. Microwaves are vital in remote sensing of ice [4]; optics for aerosols [5] and cellular components, both micrometer [6] and nanoscale [7] structures, among many other applications. In all these applications, there are cases when transmission experiments are impractical, because either the material is too opaque or, for example, has an unknown depth. The next natural choice is to use reflected, or *backscattered*, waves.

Here we ask: can one source/receiver measure the properties of a random particulate medium? And is it possible to do so without measuring the backscattering for a range of scattering angles, and without knowing the depth of the medium?

Figure 1 illustrates a backscattered wave in time measured at one point in space. We consider only elastic scattering, and scattered waves that have the same frequency as the incident wave. We show that, with this simple setup, it is possible to recover a wide range of concentrations and particle radiuses, even including particles with a sub-wavelength radius. We also identify which part of the
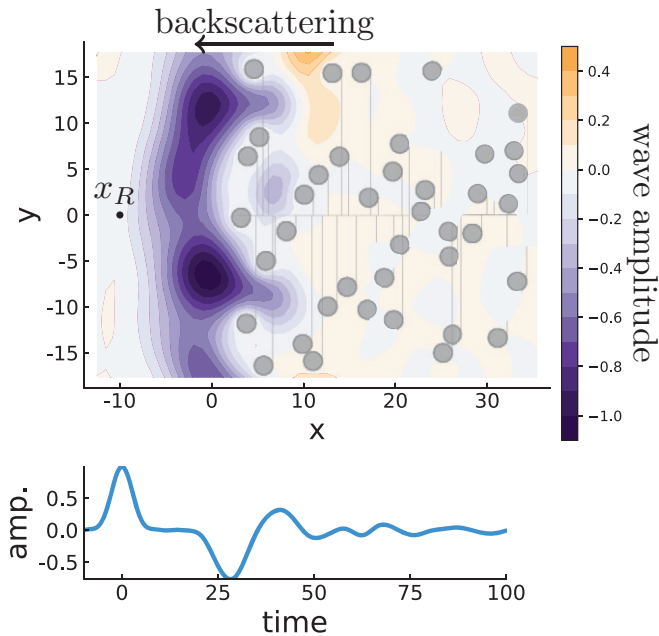
Fig. 1: (Colour online) The snapshot above is of a plane-wave pulse being backscattered by the grey particles, in the region $x > 0$, after time $t = 20$ (non-dimensional). The incident pulse originated at the line $x = x_R$, then travelled towards the particles and was then backscattered. The blue line graph shows the amplitude measured at $(x_R, 0)$ over time, where around time $t = 25$ the backscattered waves begin to arrive. The particles occupy 10% of the volume and around 150 particles were used for these simulations.

backscattered signal is sensitive to the concentration and particle radius. To achieve these goals, we use *learning curves* from supervised machine learning, and, in doing so, we also show how to accurately predict particle radius and concentration from backscattered waves. Supervised machine learning in similar contexts has already shown great promise [8,9]. See [9] for a summary of machine learning applications in remote sensing.

The long-term goal is to develop a device, as simple as possible and with little prior information, that can determine the statistical properties of the particles for a broad range of random media. To do so will require theoretical predictions, experiments and simulations of backscattered waves. A supervised learning approach can then easily combine data from these different sources to produce an algorithm that predicts particle statistics. Here we take the first step towards this goal, by using simulated data, as it is the most accurate for a broad range of media.

The most common approach to determine particle properties from backscattering, to date, is to adjust the parameters of a mathematical model until it fits the measured backscattering [2,10]. Ideally, these two approaches could be combined to produce an accurate method valid for a large range of parameters.

Simulating near-surface backscattering. – We consider a simple setting with Dirichlet boundary

conditions, that is, where the scalar wave-field $w = 0$ on the boundary of the particles, which for acoustics corresponds to zero pressure, for elasticity corresponds to zero displacement and for electromagnetism corresponds to zero electric or magnetic susceptibility, depending on the polarisation. This case is particularly challenging for many of the current theoretical approaches, as they can lead to unphysical results, even for low frequency and low concentration, as we demonstrate below. We restricted ourselves to two dimensions to lighten the computational load, which is qualitatively similar to three dimensions [11]. In the conclusion we discuss extensions to three dimensions.

Consider an incident plane wave $e^{ik(x - x_R - t)}$, where k is the wave number of the background medium, and we have non-dimensionalised by taking the phase velocity of the background to be 1. We non-dimensionalise because the theory applies to many different applications. The total wave $w = e^{ik(x - x_R - t)} + u$ satisfies the two-dimensional scalar wave equation, where u is the backscattered wave from the particles within the half-space $x > 0$ and $(x_R, 0)$ is the receiver position, such as shown in fig. 1. If the receiver is close to the particles, then near-field effects will dominate and many realisations will be needed to calculate the statistical moments. To avoid this, we choose $x_R = -10$. We use a for particle radius, $\mathfrak{n}$ for number of particles per unit area (concentration) and $\phi = a^2 \pi \mathfrak{n}$ for volume fraction, and consider a wide range of media:

$$1\% \leq \phi \leq 21\%, \quad 0.2 \leq a \leq 2.0 \quad \text{and} \quad 0 \leq k \leq 1, \quad (1)$$

for instance, these values are typically used in emulsions, suspensions, and for atmospheric aerosols.

For random media it is convenient to use the moments of u. That is, if Λ represents one configuration of particles, then $u = u(\Lambda)$ depends on Λ and its ensemble average is $\langle u \rangle = \int u(\Lambda) p(\Lambda) d\Lambda$, where $p(\Lambda)$ is the probability of the particles being in the configuration Λ, then the central moments are

$$\langle u \rangle_n = \langle (u - \langle u \rangle)^n \rangle^{1/n}. \quad (2)$$

We will now associate each medium with a fixed particle radius, concentration and set of moments $\langle u \rangle$ and (2).

There are many specialised methods to determine these moments [12–16]. Those that accurately calculate $\langle u \rangle_j$ for a broad frequency range require $\langle u \rangle$, and a common approximation of $\langle u \rangle$ is to assume that $\langle u \rangle \approx e^{ik_* x}$ inside the random media, for some effective wave number k_*. For small volume fraction ϕ and direct backscattering [17,18] this approximation leads to $\langle u \rangle \approx -i\phi(\pi a^2 k^2)^{-1} \sum_n J_n(ka)/H_n(ka) e^{i(n\pi - kx)}$, where J_n and H_n are a Bessel and Hankel function of the first kind. However, this approximation diverges when $a \to 0$, while ϕ is fixed, and leads to the unphysical result $|\langle u \rangle| > 1$. Even rigorous methods [12], deduced for moderate volume fraction, present the same problem. This problem is a result of strong scatterers, with $w = 0$ on their boundaries, completely reflecting waves at low frequencies for
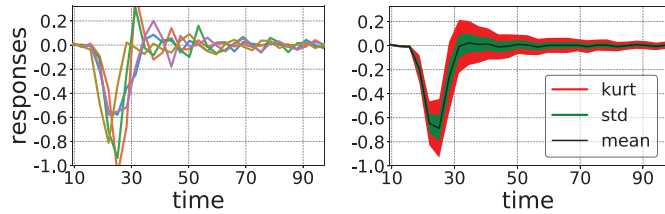
Fig. 2: (Colour online) The backscattering of the incident wave $\mathrm{e}^{-0.1(x-x_R-t)^2}$ from particles of radius $a = 0.2$ occupying $\phi = 20\%$ of the volume. Left: backscattering from five different configurations. Right: the moments of 756 configurations. The height of the black line is $\langle u \rangle$ the mean response, while the total thickness of the green and red regions are the second $\langle u \rangle_2$ (standard deviation) and fourth $\langle u \rangle_4$ (kurtosis) moment.
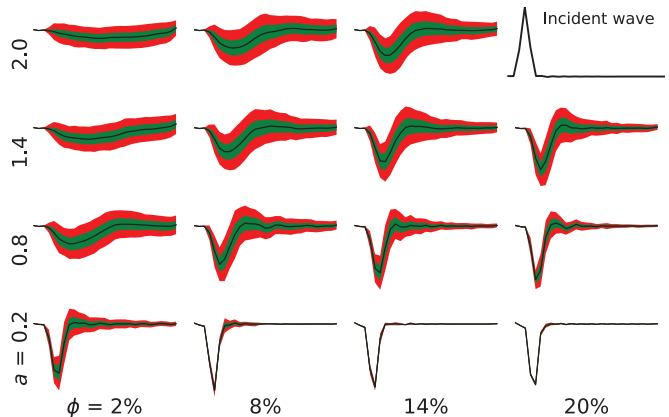

Fig. 3: (Colour online) An overview of the moments of the direct backscattering of the incident wave shown in the top right. Each graph shows $-1 < y < 0.35$, while the x-axis shows time $9.5 \leq t \leq 98$. Each column has the same volume fraction ϕ, while each row has the same particle radius a, except in the top right which is shown to the same scale as the moments, but with time $-9.5 \leq t \leq 78$ and $-0.35 < y < 1.0$.

any particle volume fraction. This can be seen by investigating the effective properties [19,20]. The consequence is that series expansions of $\langle u \rangle$ for small volume fractions do not converge for scatterers with $w = 0$ on their boundaries. For other strong scatterers with $w \approx 0$ on their boundaries, this series converges very slowly.

It may be possible to accurately describe backscattering from strong scatterers with integral methods that are valid for any volume fraction [21,22]. Though, we note, that methods derived from Lippmann–Schwinger–type equations are not formally valid for scatterers with discontinuous material properties [23], such as strong acoustic scatterers.

To accurately determine all the moments over the range (1), we use a numerical approach based on the multipole method [24] to calculate $u(\Lambda)$ for each configuration Λ, from which we determine the $\langle u \rangle_n$ with a Monte Carlo method[1]. In all our convergence tests, truncation errors and benchmarks were within 1% accuracy for each simulation. In the supplementary material `Supplementarymaterial.pdf` (SM) we explain how to reproduce our results, including high-performance software to simulate the backscattering [26] and implement the machine learning. The data used in this paper is also publicly available [27].

Approximating the backscattering from an infinite halfspace, with a limited computational domain, is challenging. To overcome this challenge we calculate the backscattering of the incident time pulse $\mathrm{e}^{-0.1(x-x_R-t)^2}$, which, for wave numbers $0 \leq k \leq 1$, results in less than 1% Gibbs phenomena, and receive the backscattering at $(x_R, 0)$. By only receiving the signal for $t \leq 98$, we can exclude from the simulation all particles that would take more than $t = 100$ for their first scattered wave to arrive at the receiver $(x_R, 0)$. That is, we need only simulate particles that are near the surface, which is why we call this *near-surface backscattering*. See fig. 2 for the incident time pulse and for $\langle u \rangle$, $\langle u \rangle_1$ and $\langle u \rangle_4$, where we include $\langle u \rangle_4$ as it is known to be sensitive the micro-structure [16,28].

In total we simulated the moments of 205 different media, evenly sampled from (1), which required 83000 backscattering simulations —each corresponding to one configuration Λ. For the larger simulations up to 7600 particles were used. To estimate the quality of the calculated moments, we used the standard error of the mean. See fig. 3 for an overview of the simulated moments.

Learning from backscattering. – With a high-quality data set of backscattered waves, we can now use supervised machine learning to generate a *model* that best fits the radius and a separate model that best fits the concentration. To test these models, we use them to predict the concentration and the radius of yet unseen media using only backscattered waves as input. Our supervised machine learning method of choice is kernel ridge regression [29,30], because when using continuous kernels it can fit any continuous function [31]. This allows us to establish whether the radius or the concentration are continuous functions of $\langle u \rangle$ or $\langle u \rangle_2$ or both. In other words, we can determine which moments are needed to predict the radius and concentration. We present the results for concentration, instead of volume fraction, because it can be accurately predicted from just $\langle u \rangle$.

Our *training* set consists of the simulated backscattered moments of 205 different media. Using this training set we *train* a model, that is to say, we use kernel ridge regression applied to the training set to generate a model. The hyperparameters of the ridge regression were selected using a 7-fold cross-validation. To determine the *predictive power* of our model, we generate a *test set* with 81 randomly chosen media with radius $0.2 \leq a \leq 2.0$ and volume fraction $1\% \leq \phi \leq 21\%$. Every medium of the test set is distinct from the training set. To measure the goodness of fit, we use the R^2 coefficient with respect to the mean of the test

---

[1]The alternative would be to piece together different theoretical methods, whose range of validity is not clear [25].
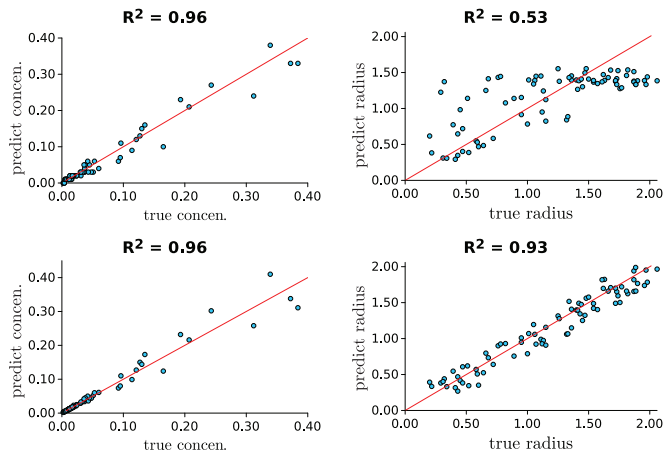
Fig. 4: (Colour online) This figure shows that to accurately predict concentration requires only $\langle u \rangle$, but to accurately predict the particle radius requires also second moment $\langle u \rangle_2$. The top two models were trained using only the mean $\langle u \rangle$, while the bottom two were trained using the mean $\langle u \rangle$ and second moment $\langle u \rangle_2$. The best prediction for the concentration gives $R^2 = 0.98$, which results from using low wave numbers, discussed later.

set. If $R^2 = 0$, then the model has the same predictive power as the mean of the test set, while $R^2 = 1$ shows that the model has perfect prediction. Finally we tested two continuous kernels, the Gaussian (or radial basis) and the Ornstein-Uhlenbeck kernel. Both kernels gave similar scores through cross-validation, though the Ornstein-Uhlenbeck kernel had a slightly better R^2 coefficient on the test set, so we only report these results.

Results. – We train two models using only $\langle u \rangle$, one to predict the concentration and one to predict the radius, see the top graphs of fig. 4. The top left and top right graphs show the scatter plot of the concentration and radius of the test set against the predicted concentration and radius, respectively. The prediction for the concentration is almost perfect, with $R^2 = 0.96$. On the other hand, the prediction for the radius is almost meaningless with $R^2 = 0.53$. The failure of the first moment $\langle u \rangle$ alone to predict the radius is significant, as it indicates that the radius is not a continuous function of $\langle u \rangle$.

To accurately predict the radius, the second moment was necessary. Indeed, training a model on the first and second moment resulted in an accurate prediction of the radius with $R^2 = 0.93$, see the bottom right panel of fig. 4.

To show that our results, such as the top right panel of fig. 4, are not due to insufficient data, and likely extend beyond our data set, we examine the learning curves. A *learning curve* shows the R^2 coefficient as the quality of the data is increased. For example, if it was possible to predict the radius from only $\langle u \rangle$, then the model's R^2 coefficient would increase when improving the training data's quality. Contrary to this, if the R^2 coefficient does not increase, or if there is no clear trend, then the model
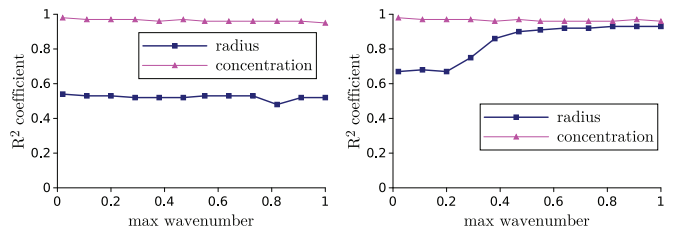

Fig. 5: (Colour online) This figure shows how increasing the maximum wave number does not lead to better predictions of particle radius when measuring $\langle u \rangle$. That is, for each point (x, y) on the graphs, we limit the incident wave numbers of the training and test set to $0 \leq k \leq x$, which results in $y = R^2$. On the left (right) we used a model trained on only $\langle u \rangle$ ($\langle u \rangle$ and $\langle u \rangle_2$).
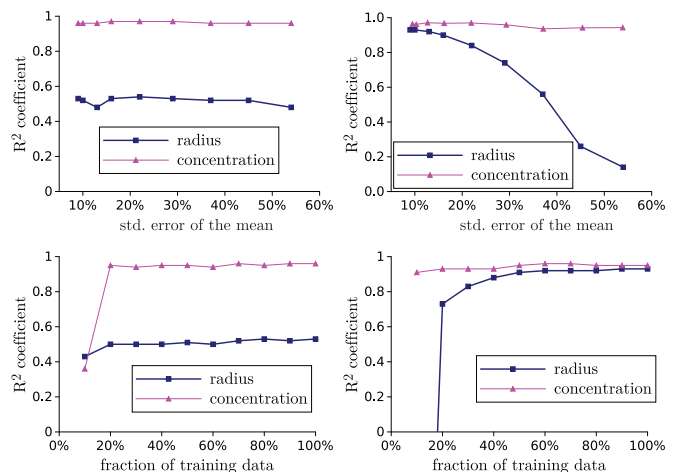

Fig. 6: (Colour online) This figure shows how well the particle radius and concentration are predicted, R^2, when changing the quality of the training data. The test set was fixed with a relative standard error of the mean of 10%. The top two graphs increase the number of simulations per medium, resulting in a change of the relative standard error of mean $\langle u \rangle$ on the x-axis. The bottom graphs increase the number of media, shown as a percentage of the full training data on the x-axis. The model on the left (right) was trained using only $\langle u \rangle$ ($\langle u \rangle$ and $\langle u \rangle_2$).

cannot predict the radius, no matter the quality of the training data.

We vary the quality of the training data by changing: the number of media, the number of simulations for each medium, and by limiting the maximum wave number k of the incident wave. For every change in the training data we re-train the model of the radius and the model of the concentration. The resulting learning curves are shown in figs. 5 and 6. The graphs on the left of all these figures are the result of using a model trained only on $\langle u \rangle$, and from them we see that the R^2 of the radius model does not tend to 1 when increasing the training data quality. The simplest explanation for this is that $\langle u \rangle$ does not by itself carry information about the radius. On the other hand, the graphs on the right of figs. 5 and 6 are models trained on $\langle u \rangle$ and $\langle u \rangle_2$, and clearly their R^2 for the radius

converges to $R^2 = 1$. In contrast, the concentration is accurately predicted from $\langle u \rangle$ even when using either 30% of the number of training media, having large standard errors of the mean, or using only wave numbers $k \leq 0.1$. In fact limiting $0 \leq k \leq 0.1$ leads to an $R^2 = 0.98$ for the concentration.

Finally, from fig. 5, we see that the learning curve saturates around a maximum wave number of 0.8. This indicates that $0 < k < 0.8$ is the ideal range to measure particles in the range $0 < a < 2$.

Conclusions. – Our results indicate that the first direct backscattered moment $\langle u \rangle$ does not carry information about a broad range of particle radiuses, for strong scatterers. However, the second moment $\langle u \rangle_2$ does carry this information. On the other hand, the particle concentration can be accurately predicted from just $\langle u \rangle$. We also demonstrated that only incident wave numbers $0 < k < 0.8$ are needed to accurately measure particles with radius $0 < a < 2$. This implies that neither theory, simulation or experiments need go beyond $ka = 1.6$, at least for strong scatterers. This also means that we are able to accurately recover radiuses that are 20 times smaller than the smallest incident wavelength.

In this study we did not consider limitations in spatial and temporal resolution, which, of course, are important in practice. However, before specialising to one particular scenario, i.e., typical acoustics and light scattering experiments, we need to know what is possible to measure or not in an ideal setting. Studies like these are therefore a vital first step. Another important step is to quantify how uncertainties in the measurements affect the prediction of the particle properties. This can be achieved by using Guassian process regression [32], which is, in a sense, a Bayesian version of kernel ridge regression.

Ultimately, our machine learning model could be embedded into a device to predict particulate properties. Though our model is initially trained on simulated data, our training procedure is simple enough that the model can be updated using real data. This step of adapting models trained on simulated data to real applications has been applied to challenging problems such as robotic grasping [33], facial recognition [34], 3D pose inference [35] and optical flow estimation [36] to name a few. These applications have advanced in strides by using simulated data, and we see a similar potential for characterising random media, such as this work.

Both the simulation (near-surface backscattering) and machine learning approach we have presented could be applied to characterise any type of particulate material from wave backscattering. To extend our approach, to 3D and other types of particles, computational efficiency is important. Simulating the backscattered moments would be faster if the multilevel Monte Carlo methods [37] and fast multipole methods [38] were used. For instance, it may be possible to measure the physical properties of the particles, as well as the size and concentration. Another avenue

to create more backscattering data is to piece together different theoretical models, which could then be validated with the numeric approach we introduced: near-surface backscattering in time.

$* * *$

REFERENCES

[1] CHALLIS R. E., POVEY M. J. W., MATHER M. L. and HOLMES A. K., *Rep. Prog. Phys.*, **68** (2005) 1541.
[2] THORNE P. D. and HURTHER D., *Cont. Shelf Res.*, **73** (2014) 97.
[3] HU P. and TURNER J. A., *J. Acoust. Soc. Am.*, **137** (2015) 321.
[4] WINEBRENNER D. P., TSANG L., WEN B. and WEST R., *IEEE J. Ocean. Eng.*, **14** (1989) 149.
[5] TORRES O., BHARTIA P. K., HERMAN J. R., AHMAD Z. and GLEASON J., *J. Geophys. Res.: Atmos.*, **103** (1998) 17099.
[6] ALMASIAN M., LEEUWEN T. G. and FABER D. J., *Sci. Rep.*, **7** (2017) 14873.
[7] YI J., RADOSEVICH A. J., ROGERS J. D., NORRIS S. C. P., APOLU L. R., TAFLOVE A. and BACKMAN V., *Opt. Express*, **21** (2013) 9043.
[8] RUPP M., TKATCHENKO A., MLLER K.-R. and VON LILIENFELD O. A., *Phys. Rev. Lett.*, **108** (2012) 058301.
[9] NOIA A. D. and HASEKAMP O. P., *Neural Networks and Support Vector Machines and Their Application to Aerosol and Cloud Remote Sensing: A Review*, in *Springer Series in Light Scattering* (Springer, Cham) 2018, pp. 279–329.
[10] WESER R., WCKEL S., WESSELY B. and HEMPEL U., *Ultrasonics*, **53** (2013) 706.
[11] GALAZ B., HAAT G., BERTI R., TAULIER N., AMMAN J.-J. and URBACH W., *J. Acoust. Soc. Am.*, **127** (2010) 148.
[12] MARTIN P. A., *J. Acoust. Soc. Am.*, **129** (2011) 1685.
[13] MISHCHENKO M. I., TRAVIS L. D. and LACIS A. A., *Multiple Scattering of Light by Particles: Radiative Transfer and Coherent Backscattering* (Cambridge University Press) 2006.
[14] SNIEDER R., *Phys. Rev. E*, **66** (2002) 046615.
[15] GARNIER J. and SLNA K., *Arch. Ration. Mech. Anal.*, **220** (2016) 37.
[16] SHENG P., *Introduction to Wave Scattering, Localization and Mesoscopic Phenomena*, in *Springer Ser. Mater. Sci.*, Vol. **88** (Springer Science & Business Media) 2006.
[17] FOLDY L. L., *Phys. Rev.*, **67** (1945) 107.
[18] GOWER A. L., SMITH M. J., PARNELL W. J. and ABRAHAMS I. D., *Proc. R. Soc. A*, **474** (2018) 20170864.
[19] PARNELL W. J. and ABRAHAMS I. D., *Waves Random Complex Media*, **20** (2010) 678.

[20] Martin P. A., Maurel A. and Parnell W. J., *J. Acoust. Soc. Am.*, **128** (2010) 571.

[21] Cobus L. A., Tiggelen B. A. v., Derode A. and Page J. H., *Eur. Phys. J. ST*, **226** (2017) 1549.

[22] Kristensson G., *J. Quant. Spectrosc. Radiat. Transfer*, **164** (2015) 97.

[23] Martin P., *SIAM J. Appl. Math.*, **64** (2003) 297.

[24] Martin P. A., *Multiple Scattering: Interaction of Time-Harmonic Waves with N Obstacles*, in *Encyclopedia of Mathematics and its Applications Series*, Vol. **107** (Cambridge University Press) 2006.

[25] Tourin A., Fink M. and Derode A., *Waves in Random Medias*, **10** (2000) R31.

[26] Gower A. L. and Deakin J., jondea/multiplescattering. jl: Version 0.1 (December 2017).

[27] Gower A. L. and Deakin J., *Backscattering from randomly placed Dirichlet particles* (version 1.0.0).

[28] Ammari H., Garnier J., Jing W., Kang H., Lim M., Slna K. and Wang H., *Mathematical and Statistical Methods for Multistatic Imaging*, *Lect. Notes Math.*, Vol. **2098** (Springer International Publishing, Cham) 2013.

[29] Boser B. E., Guyon I. M. and Vapnik V. N., *A training algorithm for optimal margin classifiers*, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT '92* (ACM, New York, NY, USA) 1992, pp. 144–152.

[30] Aizerman M. A., Braverman E. A. and Rozonoer L., *Theoretical foundations of the potential function method in pattern recognition learning*, in *Proceedings of Automation and Remote Control*, no. 25 (1964) pp. 821–837.

[31] Micchelli C. A., Xu Y. and Zhang H., *J. Mach. Learn. Res.*, **6** (2006) 2651.

[32] Rasmussen C. E. and Williams C. K., *Gaussian Process for Machine Learning* (MIT Press) 2006.

[33] Peng Xue Bin, Andrychowicz Marcin, Waremba Wojciech and Abbeel Pieter, arXiv:1710.06537 [cs.RO] (2018).

[34] Zhao J., Xiong L., Karlekar Jayashree P., Li J., Zhao F., Wang Z., Sugiri Pranata P., Shengmei Shen P., Yan S. and Feng J., *Dual-agent gans for photorealistic and identity preserving profile face synthesis*, in *Advances in Neural Information Processing Systems 30*, edited by Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S. and Garnett R. (Curran Associates, Inc.) 2017, pp. 66–76.

[35] Rad M., Oberweger M. and Lepetit V., arXiv:1712. 03904 [cs.CV] (2018).

[36] Mayer N., Ilg E., Fischer P., Hazirbas C., Cremers D., Dosovitskiy A. and Brox T., to be published in *Int. J. Comput. Vis.* (2018), https://doi.org/10.1007/s11263-018-1082-6.

[37] Giles M. B., *Acta Numer.*, **24** (2015) 259.

[38] Zhang Y. J. and Li E. P., *Progr. Electromagn. Res.*, **72** (2007) 105.