JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

Optimizing experimental design in neutron reflectometry

James H. Durant,[a] Lucas Wilkins[b]* and Joshaniel F. K. Cooper[a]*

[a]ISIS Neutron and Muon Source, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, Oxfordshire, OX11 0QX, United Kingdom, and [b]School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, United Kingdom. *Correspondence e-mail: lucas@lucaswilkins.com, lw506@sussex.ac.uk, jos.cooper@stfc.ac.uk

Received 12 August 2021
Accepted 6 April 2022

Edited by E. P. Gilbert, Australian Centre for Neutron Scattering, ANSTO, Australia

Keywords: neutron reflectivity; neutron reflectometry; experimental design; Fisher information; information theory.

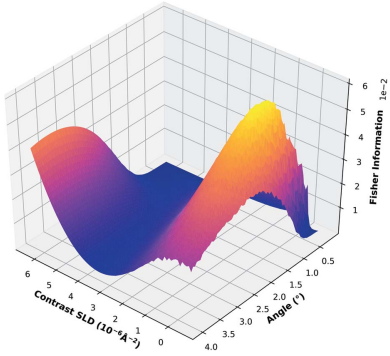
Supporting information: this article has supporting information at journals.iucr.org/j

Using the Fisher information (FI), the design of neutron reflectometry experiments can be optimized, leading to greater confidence in parameters of interest and better use of experimental time [Durant, Wilkins, Butler & Cooper (2021). *J. Appl. Cryst.* **54**, 1100–1110]. In this work, the FI is utilized in optimizing the design of a wide range of reflectometry experiments. Two lipid bilayer systems are investigated to determine the optimal choice of measurement angles and liquid contrasts, in addition to the ratio of the total counting time that should be spent measuring each condition. The reduction in parameter uncertainties with the addition of underlayers to these systems is then quantified, using the FI, and validated through the use of experiment simulation and Bayesian sampling methods. For a 'one-shot' measurement of a degrading lipid monolayer, it is shown that the common practice of measuring null-reflecting water is indeed optimal, but that the optimal measurement angle is dependent on the deuteration state of the monolayer. Finally, the framework is used to demonstrate the feasibility of measuring magnetic signals as small as 0.01 μ_B per atom in layers only 20 Å thick, given the appropriate experimental design, and that the time to reach a given level of confidence in the small magnetic moment is quantifiable.

1. Introduction

Experimental design is a deep and open-ended subject, but it is one that can often be reduced to the quantitative problem of maximizing the amount of information produced by a given experiment. It is often the case that the experimentalist has the ability to change a limited set of well defined experimental conditions, and wishes to find those conditions that will reduce the uncertainty in the outcome by the largest amount. This is a particularly pertinent problem in the field of neutron reflectometry (NR), where the analysis of reflectivity data is ill-posed and, as a consequence, the quality of the data is of great importance. This is further compounded by the significant lead time and cost of accessing a neutron beamline; a typical experiment will take months to organize, last only a few dozen hours, and have operating costs of the beamline running into the thousands of pounds per day. Despite these significant challenges, NR can provide valuable insight, for example, investigating the structural properties of lipid leaflets in biological membranes (Skoda, 2019), probing magnetism in thin-film heterostructures (Liu & Ke, 2015), examining surface chemistry at the air/water interface (Welbourn & Clarke, 2019) and exploring layer structures in organic photovoltaics (Zhang *et al.*, 2017).

Typically, the design of an NR experiment is determined by the knowledge gained in similar previous experiments, *i.e.* trial and error and 'rules of thumb'. To provide more rigorous and



quantifiably optimized designs, we developed a framework (Durant *et al.*, 2021*a*) using the Fisher information (FI) (Fisher, 1925) that quantifies the maximum information in an NR experiment using the assumption of Poisson-based counting statistics. Using the FI and accurate experiment simulation, the framework was shown to be capable of optimizing experimental design in a computationally inexpensive manner. In this work, we extend the framework and demonstrate its utility in experimental design of a diverse selection of commonly measured systems.

In NR, a collimated neutron beam is directed onto a surface of a sample and the intensity of the reflected radiation is measured as a function of angle and neutron wavelength, and for polarized measurements, spin state. Therefore, as an experimenter, one is presented with the non-trivial choice of angle(s) and counting time(s) for any NR measurement. The resulting reflectivity profile provides insight into the structure and properties of the sample surface, including the thickness, scattering length density (SLD, the product of a material's density and its neutron scattering length) and interfacial roughness of any thin films. For certain systems, additional properties can be inferred, such as the sample hydration in liquid-based experiments and magnetic moments in magnetic samples; these systems often present added complexity in their experimental design, *e.g.* contrast choice.

In NR analysis, one is usually tasked with inversion of the reflectivity curve, with the aim of reconstructing the SLD profile. This is a known inverse problem due to the loss of phase information upon scattering (Majkrzak & Berk, 1995) and, consequently, the analysis is predominantly model dependent. An initial model is typically defined using a series of contiguous layers and is informed by prior knowledge of a system and the underlying science. The model reflectivity is calculated using the Abelès matrix formalism for stratified media (Abelès, 1948) or the Parratt recursive method (Parratt, 1954). The difference between the model reflectivity and the data is then calculated, using a metric such as the chi-squared, and the model is then iteratively modified in order to find the best agreement between the two. For this work, we use the Python packages *refnx* (Nelson & Prescott, 2019) and *Refl1D* (Kienzle *et al.*, 2017) for our model definitions and reflectivity calculations. Since it is an inverse problem, NR analysis is provably unsolvable in the general case with no additional information. Approximation methods have been devised to alleviate this issue and find likely solutions, including the use of evolutionary algorithms (Storn & Price, 1997) and even neural networks (Mironov *et al.*, 2021; Loaiza & Raza, 2021; Doucet *et al.*, 2021; Greco *et al.*, 2021). However, there is no guarantee that such methods will provide the sample properties describing the 'true' SLD profile. For a given reflectivity curve, there will be a large number of models which give equivalently good fits to the data. As a result, it is imperative that an experiment is well designed to produce data that support the experimentally measured sample as best as possible.

For advanced data analysis, a Bayesian approach is commonly taken, often employing the use of sampling methods such as Metropolis–Hastings Markov chain Monte Carlo (MCMC) (Metropolis *et al.*, 1953; Hastings, 1970) and nested sampling (Skilling, 2004, 2006). These methods approximate parameter values and confidence bounds through estimation of the parameter posterior distributions. For this work, we use nested sampling implemented in the Python package *dynesty* (Speagle, 2019). In general, sampling methods will not be able to extract the maximum information that a data set contains, as calculated using the FI. They instead determine the maximum extractable information, given the limitations of the data, *e.g.* parameter correlations which limit the ability to determine certain values. We use the computationally expensive sampling methods as a form of 'ground truth' to validate the improvements in individual parameter variances and parameter-pair covariances using the experimental designs determined from the FI.

By simulating experiments with a known model, the FI framework can be used to calculate the information content of any set of experimental conditions without requiring expensive fitting, sampling calculations or beamtime for data acquisition. The experiment simulation has been shown to be accurate, fast and general; any beamline and reflectometer can be simulated, given the instrument-specific incident flux profile. In our previous work, we used the framework to consider the experimental design of a common model system for structural biology: a 1,2-dimyristoyl-*sn*-*glycero*-3-phosphocholine (DMPC) bilayer deposited onto a silicon surface. In this work, we apply a new optimization approach, using the FI, to this system and several others that are frequently measured from different fields. We strive to answer some of the fundamental questions presented in the design of NR experiments, including the choice of measurement angle(s), counting time(s), bulk water SLD(s) in liquid-submerged samples, and underlayer SLD(s) and thickness(es) in lipid bilayer and magnetic samples. These example model systems have been chosen to best demonstrate the wide applicability of the framework and can be modified to investigate other experimental conditions of interest. All examples shown here are presented on our GitHub repository (Durant *et al.*, 2021*b*), alongside the underlying code for more advanced investigations.

2. Methods

2.1. Reflectometry models

In NR, a model is defined by a structure representing a physical sample, some level of background noise, an experimental scale factor and the instrument resolution function. The way this structure is defined is unimportant in our framework, as long as the model reflectivity can be calculated at a given neutron momentum transfer, Q,

$$Q = \frac{4\pi \sin\theta}{\lambda}, \tag{1}$$

where λ is the neutron wavelength and θ is the measurement angle. For this work, our model structures consist of contiguous

layers of homogeneous SLD (*i.e.* slab models). For our experiment simulation, we use the Poisson-based counting statistics approach shown alongside the FI derivation in our previous work (Durant *et al.*, 2021*a*). As input, this method simply requires a model (or multiple models if multiple data sets are required) and the incident flux as a function of wavelength for the instrument being simulated. For this work, the flux profile (both polarized and non-polarized) was taken on the OFFSPEC neutron reflectometer (Dalgliesh *et al.*, 2011). This flux profile is used with the model(s) to calculate the expected number of neutrons in each Q bin, as is required for the FI. The framework also requires parameter values to be input which can be estimated from the given data or specified manually if known. We use *refnx* and *Refl1D* to define our models and to load the models' associated measured or simulated reflectivity data.

The simulation framework was developed on the principle that, for a given measurement angle, each neutron wavelength corresponds to a single momentum-transfer point, Q, as is often the case in reflectometry. As a result, the experiment simulation would require modification for other techniques where this is not the case, *e.g.* small-angle neutron scattering. Further, the simulation was developed to synthesize data without background subtraction.

2.2. Beamline setup

A conventional neutron reflectometer relies on two sets of apertures in order to collimate a neutron beam, limiting the beam footprint at the sample point. For the majority of the examples shown in this work it is appropriate to limit the beam footprint to be less than the sample extent, *i.e.* under-illuminating, so as not to add background noise by illuminating sealing gaskets or similar. When using an algorithm such as that described by Cubitt *et al.* (2015), the angular resolution of the measurement is dependent on the sample-to-detector distance and the detector pixel size. Therefore, there is an infinite set of aperture openings with equivalent resolutions that will achieve the same sample footprint. We chose the openings that maximize the flux on the sample, which can be proven to be half the maximum opening (which stays within the footprint) for each aperture. Thus, the beamlines here are already optimized (in hardware) for maximum flux, *i.e.* maximum information gain. For systems where this algorithm cannot be applied, the effect of resolution on the model (and therefore information) may need to be considered.

In this work, we consider the beamlines at the ISIS Neutron and Muon Source at the Rutherford Appleton Laboratory. These beamlines very commonly operate with a footprint of ∼60 mm, as much of the sample environment equipment has been standardized. For the non-polarized experiments described in this work, we assume a footprint of 60 mm (polarized experiments are described below). For all simulations, both apertures are scaled linearly with measurement angle to maintain a constant footprint on the sample. This scales the incident flux by a factor proportional to the angle squared and assumes no changes in beam profile with opening apertures; this is generally true in practice. Thus, when simulating an experiment, only a single incident flux profile is required, and the profile is scaled appropriately to the measurement time and angle.

It is not common practice on the ISIS beamlines to apply background subtraction, the preference normally being to fit the background in the analysis software. This is not the case in all reflectometer beamlines and many rely on background subtraction for all of their measurements, so we mention some considerations which should be taken into account when using this framework. The issue lies in the fact that the statistics governing background-subtracted data are no longer Poisson distributed but follow something like a Skellam distribution (*i.e.* the difference between two Poisson distributions). In this case our derivation of the FI no longer holds since equation (4) of Durant *et al.* (2021*a*) is not correct for anything other than a Poisson distribution. The best way to circumvent this issue is to include the background in the model as a parameter; our framework then still applies, and indeed from a statistics point of view, it is a more correct procedure. From a practical standpoint, we believe that given the uncertainties associated with model parameters before they are measured (*i.e.* in the experimental design phase) a sufficiently optimal experiment may still be designed solely using a modification to the incident flux profile and its behaviour at higher Q. Irrespective, the treatment of the data post-experiment should not affect any experimental design as long as an appropriate background term is included in the model.

2.3. Experimental design optimization

The derivation and initial demonstration of the FI for NR are given fully by Durant *et al.* (2021*a*) and will not be repeated here. Instead, we focus on how the systems being investigated are described and how the FI is utilized in optimizing their experimental design.

By refining the ideas introduced in our initial work, we have developed an improved methodology for experiment optimization using the FI. Previously, we had not been able to directly compare the information content of parameters of differing units, since parameters can be on widely differing scales and the units of the FI are 'nats' (natural units of information) per parameter unit squared. We addressed this issue by weighting parameter importance. We have assumed that our model parameters are all of equal importance in terms of their specified units, but the methodology can be easily adapted to accommodate a custom importance weighting if desired. Another limitation of our previous work was that we had only considered the diagonal entries of the FI matrix (the information content of each individual parameter). With all parameters now on the same scale, we can include the off-diagonal elements of the FI matrix by finding the experimental conditions that maximize the minimum eigenvalue: a maximin approach. This effectively finds the conditions that improve the 'worst' possible combination of parameters the most, thereby minimizing the overall uncertainty of the experiment. The reader is referred to the

supporting information (SI) for further details of this approach and the importance scaling.

One benefit of the new optimization method is that the optimization function is multivariate scalar, whereas our previous approach required optimization of a vector of parameter information values. Bounds are placed on the variable experimental conditions and so the optimization is constrained. It is therefore possible to take every experimental condition into account simultaneously and find the 'complete' optimal experimental setup using established optimization algorithms, *e.g.* differential evolution (DE) (Storn & Price, 1997); for a minimizing algorithm like DE, the negative of the minimum eigenvalue can be used as the optimization objective. This 'complete' optimization can account for any complex relationships between experimental conditions that may be missed if these conditions were considered individually.

2.4. Lipid bilayers

In this work, we take two lipid bilayer models and investigate the choice of measurement angles, counting times, contrasts and the effect of underlayers on the models' parameter uncertainties. The first of these models is taken from our previous work where we quantified the information content of each model parameter of a DMPC bilayer deposited onto a silicon surface as a function of the bulk water SLD. The bilayer model was defined by two lipid leaflets with fixed surface coverage. The lipids were measured against two water contrasts, H_2O and D_2O, using the CRISP neutron reflectometer (Penfold *et al.*, 1987) as part of the ISIS neutron training course. The data were simultaneously fitted using *RasCAL* (Hughes, 2017) with the fitting constrained against measured data for a bare Si/D_2O interface including a native SiO_2 layer. The fitted model was converted to *refnx* and reparameterized as a function of the bulk water contrast SLD, enabling simulation of the bilayer on the OFFSPEC reflectometer with arbitrary contrast SLD. Further details of the model parameterization can be found in our previous work (Durant *et al.*, 2021*a*).

In addition to the DMPC bilayer, we consider a more sophisticated bilayer model in this work. We take a data set from Clifton *et al.* (2016) who studied a highly asymmetric bilayer structure made of a phospholipid-rich inner leaflet composed of 1,2-dipalmitoyl-*sn-glycero*-3-phosphocholine (DPPC) and a Ra lipopolysaccharide (LPS) outer leaflet. Specifically, we investigate tail-deuterated DPPC (d-DPPC) data. Three isotopic contrasts were measured for the DPPC/RaLPS bilayer using the INTER reflectometer (Webster *et al.*, 2006): d-DPPC/RaLPS in D_2O, d-DPPC/RaLPS in silicon-matched water (SMW) and d-DPPC/RaLPS in H_2O. As with the DMPC bilayer, the model for this data was originally defined and fitted with *RasCAL* but was recreated in *refnx* for this work. The model accounted for each individual layer hydration, the presence of bilayer defects across the surface and the lipid asymmetry. For additional details of the model parameterization, the reader is directed to the SI where the fitted model parameters for both models can also be found.

Like in our previous work, we have assumed that the molecular volumes do not vary with measurement conditions. These volumes may not necessarily be constant in practice (Campbell *et al.*, 2018), but to simplify the models, the volumes have remained fixed. Using this assumption, we can optimize the choice of contrast(s) for both bilayer models. To support the addition of underlayers, whose presence changes the reflectivity curves and may improve the information gain, we set the SiO_2 hydration to 0%. We then add a layer of given SLD and thickness between the SiO_2 and the inner bilayer headgroups for each underlayer added; a fixed 2 Å roughness and 0% hydration are assumed for each underlayer. With the addition of these underlayers, the SiO_2 hydration and SiO_2/bilayer roughness parameters are excluded from the FI calculation.

When simulating data for both bilayer models, angles of 0.7 and 2.3°, and times of 15 and 60 min, respectively, are used (or the same ratio of times between angles is used when the total time is altered) with 100 data points per angle. The experimental scale factor and instrument resolution function used are 1.0, 5×10^{-6} and constant 2% dQ/Q, respectively. The background level varied between 2×10^{-6}, for D_2O, and 4×10^{-6}, for H_2O. We optimize the measurement angles over the interval $[0.2, 4.0]^\circ$, representing realistic bounds on the physically possible measurement angles of the OFFSPEC reflectometer being simulated. The contrast SLDs, underlayer thicknesses and underlayer SLDs are optimized over the intervals $[-0.56, 6.36] \times 10^{-6}$ Å^{-2} (pure H_2O to pure D_2O), [0, 500] Å and $[1, 9] \times 10^{-6}$ Å^{-2}, respectively, the latter being the range of SLDs within which most non-isotopically enriched film materials would fall.

We first consider a simple form of optimization where we visualize the optimization space for the simultaneous choice of two contrasts for both bilayers, assuming no prior measurements and assuming that the two contrasts are measured for equal amounts of time. We then validate the improvement attained with the optimal solution, suggested by the framework, using nested sampling. Following this, we generalize by optimizing the choice of up to four contrasts and additionally optimize the proportion of the total time spent measuring each contrast. However, as we go beyond optimizing two experimental conditions at once (*e.g.* three or more contrasts), we run into two problems: a brute-force approach becomes computationally difficult or infeasible, and the optimization space becomes challenging to visualize. Therefore, for our more advanced optimization results, we apply the DE algorithm. To significantly reduce the dimensionality and complexity of the optimization problem, we make the simplifying assumption that all contrasts are measured using the same angles and the same proportion of times between angles (but the total time spent measuring each contrast is not necessarily equal). In practice, this may not always be the case. However, for these model systems, it is unlikely that the added complexity of considering different angles (and counting time splits between angles) for each contrast would influence the end results drastically. Using the simplifying assumption, we optimize contrasts and then separately optimize the

measurement angles (and proportion of the time spent measuring each angle) for the contrasts.

Thus far, we have considered the choice of measurement angles, counting time ratios and contrasts for the bilayer models but we also have the ability to improve the experiment through modification of the sample structure itself. When a chemical surface other than silicon dioxide is required to interact with lipids, it is common practice to use a gold layer grown onto the silicon. This can be functionalized almost arbitrarily using self-assembled thiol-based monolayers to give any desired chemical termination. In addition to chemical functionalization there is also a portion of the field using magnetic underlayers as an additional contrast mechanism. Commonly, a permalloy or nickel thin film is deposited onto the substrate and then coated in gold again with a functionalized surface. The whole system is then measured with a polarized neutron beam, using the different SLDs the two spin states experience in the permalloy either instead of changing the water contrast or in addition to it. Such systems offer a wealth of potential optimization; however, the complexity of them means that the benefits of the magnetic contrasts will be highly model, and even beamline, specific. Thus, we do not investigate them further in this work, but we do expect the tools described here to have huge utility in designing such experiments. We instead demonstrate the ability to increase experimental information solely through the addition of an underlayer or underlayers whose SLDs and thicknesses are optimized using DE. We also visualize the optimization space for a single added underlayer, compare the results with gold and permalloy underlayers in common use (Clifton *et al.*, 2015), and use nested sampling to validate the improvement in the estimated parameter uncertainties and posterior distributions. We note that, when investigating the parameter uncertainties, we do not include the underlayer parameters, since they are essentially 'nuisance' parameters, whose values are needed but not important for the comparison. In real experiments the underlayer parameters are likely to be well defined enough by the initial characterization of the wafer (due to strong scattering and a large number of fringes) that very little additional ambiguity will be included in the bilayer parameters.

When competing experimental conditions are involved, we perform our optimization using DE and assuming a fixed time budget; when optimizing contrasts (or angles), the time spent measuring all contrasts (or angles) is constrained to be constant and the proportion of the total time spent measuring each contrast (or angle) is optimized in addition to the contrast SLDs (or angles) themselves. The optimization is also performed using a large total time budget to reduce the impact of noise in the simulated data and assist in convergence to the global optimum. To illustrate this, suppose we were optimizing two contrasts and DE was in a region of the optimization space where one contrast had a counting time of 1% and the other 99% of the total time budget, and the total time budget was only 10 min; the former contrast would only have a counting time of 6 s, resulting in very poor data and thus hindering the optimization process. Note that using a large total time does not influence the results of optimization (other than reducing the effect of noise); if the total time is increased and the ratio of times between contrasts remains fixed, there is no change in the relative difference in information between contrasts.

2.5. Kinetics

Often in NR, we wish to make measurements where the sample changes over time, for example, where a sample degrades due to some reaction. Typically, when measuring these kinetic systems, only a single measurement is possible, and the designs of such experiments must reflect this. Therefore, when considering such a design, a time budget becomes irrelevant as there are no longer multiple conditions to consider splitting a time budget between. Instead, we wish to make the measurement that results in the greatest information about the system using the experimental conditions we do have control over; for a liquid-submerged sample, these are the measurement angle and liquid contrast.

To demonstrate the framework's flexibility in the experimental design of a kinetic system, we modelled a system (Clifton *et al.*, 2011) that investigated the binding of puroindoline-a (Pin-a) proteins to lipid monolayers composed of 1,2-dipalmitoyl-sn-*glycero*-3-phospho-(1-rac-glycerol) (DPPG); the original measurements were carried out using the SURF neutron reflectometer (Penfold *et al.*, 1997) with the isotopic contrasts measured being an equilibrium Pin-a adsorbed d-DPPG (chain-deuterated DPPG) monolayer on null-reflecting water (NRW), h-DPPG (hydrogenated DPPG) on NRW and h-DPPG on D_2O. The model was originally defined and fitted by *RasCAL*, but for this work we have recreated the model in *refnx* without a protein, leaving just a monolayer. In the original work, the model SLDs and thicknesses were fitted directly and then converted to volume fractions. We use a more sophisticated model here, defined by area per molecule (APM). The model can describe both h-DPPG and d-DPPG, accounting for the headgroups containing water through defects across their surfaces and also the water bound to the hydrophilic headgroups. The model has been parameterized as a function of the bulk water contrast SLD to facilitate contrast optimization. The full details of the model parameterization and fitting can be found in the SI. As with Section 2.4, we have assumed the molecular volumes are invariant with changing experimental conditions.

To introduce kinetics into the model, we simulate the surface excess decreasing over time by increasing the lipid APM, since the surface excess (in mg m^{-2}), Γ, is related to the APM (in $Å^2$) by

$$\Gamma = \frac{\text{Molecular weight } \times 10^{23}}{N_A \times \text{ APM}}. \quad (2)$$

$N_A$is Avogadro's constant. Typically, the reflectivity profile of a monolayer is fairly featureless and so the model parameters are not particularly well defined when only considering a single contrast (as is the case for our model system). As a consequence, we only investigate the parameter of greatest

significance: the lipid APM. As there is just a single parameter, there are no parameter-pair covariances to consider and so we can simply use the FI itself as our objective to maximize.

For our results, we simulate data using an experimental scale factor of 1.0, instrument background of 5×10^{-6} and resolution function of constant 2% dQ/Q. To account for the sample degradation, data are simulated for 20 lipid APM values, ranging from the fitted value of 54.1 to 500 Å^2, with the FI calculated over the entire simulated data set; this quantifies the maximum information obtainable about the APM over the full experiment using the given conditions. The entire data set is given a time budget of 150 min with 100 data points per APM value. When optimizing the experimental design, the measurement angle and contrast SLD are optimized over the intervals $[0.2, 4.0]^\circ$ and $[-0.56, 6.36] \times 10^{-6}$ Å^{-2}, respectively.

2.6. Magnetism

An area where it is particularly important to be able to discern small signals and maximize differences between models is in thin-film magnetism. It is often the case that small moments are induced by a ferromagnet in neighbouring layers, proximity magnetism (Khaydukov *et al.*, 2013; Cooper *et al.*, 2017; Duffy *et al.*, 2019; Zhan *et al.*, 2019; Inyang *et al.*, 2019), or a layer might only have a very small moment in the first place, but it is particularly important to know the magnitude of this moment. The advantage of neutrons in these situations is that they provide an absolute moment, so in many ways are the idea tool for the task. However, neutron sensitivity is much lower than the sensitivity of some other techniques, with a rough rule of thumb putting the minimum measurable moment with polarized neutron reflectivity at around 0.05 μ_B per atom.

In this work, we use a data set (Cooper *et al.*, 2017) that was measured to quantify the moment that an yttrium iron garnet (YIG) film, grown on an yttrium aluminium garnet (YAG) substrate, induces in an adjacent platinum capping layer. This experiment was already unusually sensitive to the induced moment, allowing a limit of $\pm 0.02\ \mu_B$ per atom to be placed on the induced magnetism. The model for this work was originally defined and fitted in the reflectivity analysis package *GenX* (Björck & Andersson, 2007) but was recreated in *Refl1D* for our experimental design analysis. The polarized, but not analysed, measurements were made using the MAGREF neutron reflectometer (Ambaye *et al.*, 2008). The reader is directed to the SI for further details of the model parameterization and fitting.

To keep the focus of the experimental design on the sample itself, we use three pre-defined incident angles of 0.5, 1.0 and 2.0° with measurement times of 30, 60 and 120 min, respectively (or the same ratio of times between angles when the total time is altered), and 100 data points per angle: a commonly used array of angles and times for OFFSPEC when polarized. More angles are required than for unpolarized measurements since polarizing systems reduce the incident flux at low wavelengths. As such, a different incident flux profile for OFFSPEC is used for this section, suitable for simulating polarized data. For our simulation, we use an experimental scale factor, level of background noise and instrument resolution function of 1.0, 5×10^{-7} and constant 2% dQ/Q, respectively. We assume a scale factor of unity, *i.e.* every incident neutron interacts with the sample, but this is usually not the case for magnetic samples due to the difficulties of getting large homogeneous films. For scale factors less than unity, the results shown here will hold true, but the counting times will scale inversely to the scale factor (*e.g.* a scale factor of 0.1 would require counting 10 times longer to achieve the same uncertainty).

Without changing any of the interface physics, we are able to reparameterize the model as a function of the thicknesses of both the YIG and platinum layers in a quest to make our experiment more sensitive to the moment. For simplicity we assume that there is a constant moment of 0.01 μ_B per atom (equivalent to a magnetic SLD of 0.0164×10^{-6} Å^{-2}) in the 21 Å-thick platinum layer, *i.e.* within our current error bound. When changing the platinum layer thickness, we constrain the magnetic moment to remain within this 21 Å layer so as not to increase the total magnetism in the system as the platinum layer thickness increases. We calculate the optimal thicknesses of the YIG and platinum layers in the intervals [400, 900] Å and [20, 100] Å, respectively.

To validate the improvement with the suggested experimental design, we use the ratio of likelihoods between two models, one with an induced moment of 0.01 μ_B per atom in the platinum layer and one with no moment, as a function of measurement time, to determine what level of statistics is required for a differentiable difference between the two models. The likelihood, $\mathcal{L}$, provides a measure of difference between a given data set and model and, for this work, is defined as

$$\ln \mathcal{L} = -\frac{1}{2}\sum_{i=1}^{N}\left\{\left(\frac{r_i - r_{i_m}}{\delta r_i}\right)^2 + \ln\left[2\pi(\delta r_i)^2\right]\right\}, \qquad (3)$$

where N is the number of data points, r_i is the experimental/simulated reflectivity at the ith Q point, δr_i is the uncertainty in the experimental/simulated reflectivity at the ith Q point and r_{i_m} is the model reflectivity at the ith Q point calculated using the Abelès matrix formalism. The data for the models are simulated with the induced moment and hence we should expect the correct model to become more 'likely' as the measurement time is increased. This process is performed twice, once with an optimized design and once with a sub-optimal design, to illustrate how the optimized design reduces the time to confidently discern the magnetic moment.

3. Results and discussion

3.1. Lipid bilayers

3.1.1. Contrasts. The fitted SLD profiles and experimental reflectivity data for the two lipid bilayer systems of Section 2.4 are shown in Fig. 1. Fig. 1 also visualizes the optimization space for the simultaneous choice of two contrasts, for the two models, assuming no prior measurement. The plot has

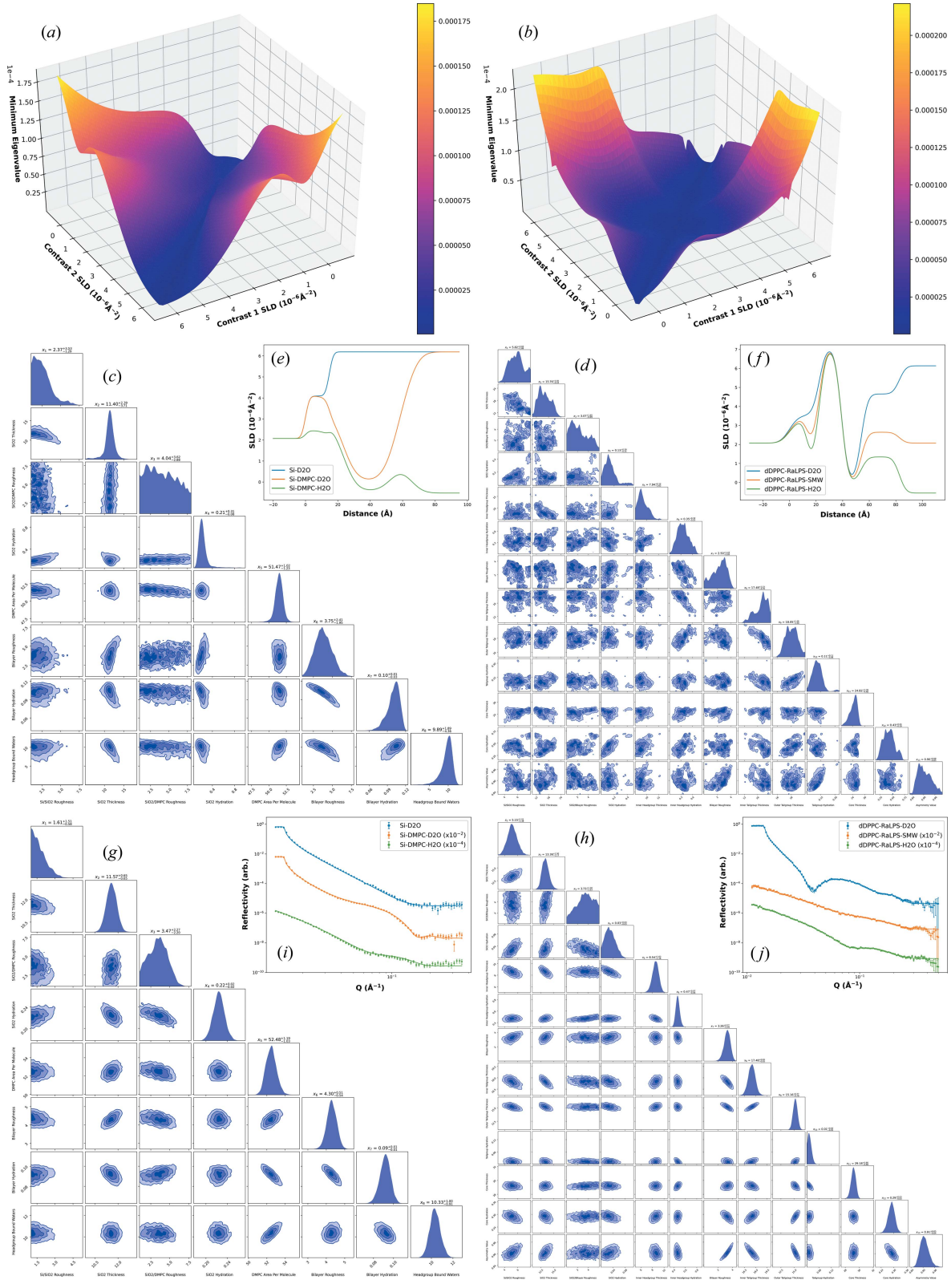


Figure 1
Plots on the left correspond to the DMPC bilayer and plots on the right to the DPPC/RaLPS bilayer. Shown [(*a*) and (*b*)] are the plots of minimum eigenvalue versus bulk water contrast SLD for the bilayer models, for the simultaneous choice of two contrasts assuming no prior measurement. Higher values of minimum eigenvalue are better. Also shown are the nested sampling corner plots from sampling simulated data of solely D_2O [(*c*) and (*d*)] and simulated data of D_2O and H_2O [(*g*) and (*h*)]. The insets are the fitted SLD profiles [(*e*) and (*f*)] and reflectivity curves [(*i*) and (*j*)] for the experimentally measured data sets of the two models. For clarity, the H_2O and D_2O DMPC bilayer data sets have been offset by factors of 10^{-2} and 10^{-4}, respectively. The H_2O and SMW DPPC/RaLPS data sets have been offset by the same factors, respectively. All figures are generated by scripts available at GitHub (Durant *et al.*, 2021*b*), with full-resolution figures stored there also.

symmetry as, assuming that each contrast is measured under the same conditions, the choice of contrast is commutative (the order does not matter). For both bilayer models, it is clear that measuring contrasts of maximum difference in SLD is optimal and that measuring the same contrast twice (*i.e.* for twice as long) or contrast matching a layer are both sub-optimal. These results agree with both common practice and sampling methods. This is illustrated in Fig. 1 where the nested sampling corner plots (using the default *dynesty* stopping criteria) from sampling simulated data of D_2O and H_2O, and D_2O measured for twice as long, are shown. The distributions are clearly much better defined (*i.e.* more Gaussian) in the D_2O and H_2O case when compared with only D_2O. In addition, the 95%/2σ credible intervals of the parameters are considerably lower in the former case when compared with the latter, with the median parameter error at ∼50% of the parameter value for only D_2O, decreasing to ∼10% with the addition of the second contrast.

3.1.2. Time-constrained optimization. Table 1 shows the optimized contrasts and counting time splits for each bilayer model using one to four contrasts, assuming a fixed time budget. For a single contrast, it appears that D_2O is optimal for DMPC and H_2O optimal for DPPC/RaLPS. Since the DMPC bilayer tailgroup is hydrogenated and the DPPC tailgroup deuterated, we hypothesize that the contrasts are optimal as a result of maximum contrast variation. For both models, there is a clear and large improvement in measuring two contrasts over just one. This is particularly noticeable for the DPPC/RaLPS model as it is defined using a large number of model parameters which are poorly described with just a single contrast. When measuring two contrasts, the results clearly agree with those of Fig. 1: measuring D_2O and H_2O is optimal for both models, given the measurement conditions detailed in Section 2.4.

When considering three contrasts, it can be seen that a small improvement can be obtained with an additional contrast of SLD around 1.9×10^{-6} Å^{-2} for DMPC and 3.4×10^{-6} Å^{-2} for DPPC/RaLPS. This conforms with current practices where D_2O and H_2O are typically measured and a third contrast of SMW (SLD of 2.07×10^{-6} Å^{-2}) is sometimes also measured. Although these contrast SLDs differ from SMW, an improvement can still be attained from measuring SMW. This is illustrated in Fig. 2 which shows how the minimum eigenvalue changes with a third contrast SLD for both models; the split of the total counting time between each contrast was defined by the three contrast results of Table 1. We emphasize that the minimum eigenvalues should not be compared between models (*e.g.* DMPC versus DPPC/RaLPS), since they have a different number of parameters, but rather the water contrast at which they are individually maximized is of importance. When considering which third contrast improves both models the most, SMW is an ideal candidate and probably a suitable choice if the optimal (model-dependent) third contrast is unknown. The gains achieved in measuring a third contrast are relatively small and the time to change contrast may outweigh the minor gains achieved when compared with measuring either D_2O or H_2O for longer. Therefore, we proceed with the assumption that D_2O and H_2O will be measured. Finally, the results suggest that measuring more than three contrasts is unnecessary; given four or more contrasts, the solution is essentially the same as with three after combining the times of (near) identical contrast SLDs.

Table 1
Optimized contrast SLDs and counting time splits for the DMPC and DPPC/RaLPS bilayer models using one to four contrasts.

Also shown is the minimum eigenvalue for each set of conditions which was maximized using DE optimization.

	Contrast SLD (10^{-6}Å^{-2})				Split of time (%)				
Sample	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Minimum eigenvalue
DMPC	6.36	–	–	–	100.0	–	–	–	4.957×10^{-6}
	−0.56	6.36	–	–	70.7	29.3	–	–	1.919×10^{-3}
	−0.56	1.91	6.36	–	22.0	26.5	51.5	–	2.666×10^{-3}
	−0.56	1.90	1.93	6.36	22.7	5.8	19.4	52.1	2.666×10^{-3}
DPPC/RaLPS	−0.56	–	–	–	100.0	–	–	–	6.367×10^{-8}
	−0.56	6.36	–	–	21.9	78.1	–	–	2.771×10^{-3}
	−0.56	3.38	6.36	–	24.5	21.5	54.0	–	4.387×10^{-3}
	−0.56	3.37	3.41	6.36	25.0	3.8	17.7	53.6	4.387×10^{-3}

From Table 1 it can also be seen that the time spent measuring each contrast is highly model dependent; for the DMPC model, a much smaller proportion of the total time budget needs to be spent measuring D_2O when compared with the DPPC/RaLPS model. This may initially seem surprising when referring to our previous hypothesis that the maximum information is obtainable in the contrast opposite to that of the tailgroup, *e.g.* D_2O for hydrogenated tailgroups. However, we need to consider that there will be important information in both D_2O and H_2O contrasts, but it will take more time to obtain the equivalent complementary information in the contrast matching the tailgroup deuteration state when compared with the contrast with maximum variation. We see that this is indeed the case when looking at the optimal time splits for two contrasts; the contrast with the most information (*i.e.* that which is chosen when only a single contrast can be

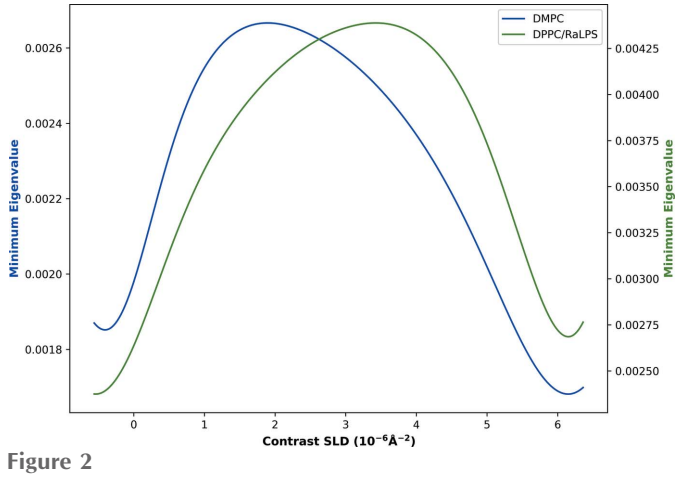


Figure 2
Minimum eigenvalue versus bulk water contrast SLD for the DMPC (blue) and DPPC/RaLPS (green) bilayer models, for a third contrast choice, assuming D_2O and H_2O have been previously measured.

Table 2
Optimized measurement angles and counting time splits for the DMPC and DPPC/RaLPS bilayer models using one to four angles.

Also shown is the minimum eigenvalue for each set of conditions which was maximized using DE optimization.

	Angle (°)				Split of time (%)				
Sample	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Minimum eigenvalue
DMPC	3.17	–	–	–	100.0	–	–	–	1.027×10^{-2}
	0.82	4.00	–	–	11.9	88.1	–	–	1.599×10^{-2}
	0.8	3.94	4.00	–	11.8	2.9	85.3	–	1.599×10^{-2}
	0.8	3.98	3.99	4.00	11.6	45.0	32.1	11.3	1.599×10^{-2}
DPPC/	2.91	–	–	–	100.0	–	–	–	1.035×10^{-2}
RaLPS	0.68	4.00	–	–	8.9	91.1	–	–	2.314×10^{-2}
	0.68	4.00	4.00	–	9.1	5.8	85.2	–	2.313×10^{-2}
	0.43	0.69	4.00	4.00	0.0	8.7	61.4	29.9	2.313×10^{-2}

Table 3
Optimized underlayer SLDs and thicknesses for the DMPC and DPPC/RaLPS bilayer models using zero to three underlayers.

Also shown is the minimum eigenvalue for each set of underlayer properties which was maximized using DE optimization. Layer 1 is nearest to the silicon substrate and layer 3 is nearest to the water solution.

	Underlayer SLD (10^{-6} Å^{-2})			Underlayer thickness (Å)			
Sample	Layer 1	Layer 2	Layer 3	Layer 1	Layer 2	Layer 3	Minimum eigenvalue
DMPC	–	–	–	–	–	–	4.694×10^{-4}
	5.39	–	–	127.1	–	–	4.016×10^{-3}
	5.19	8.99	–	50.2	105.9	–	5.284×10^{-3}
	2.32	5.34	8.99	29.3	46.8	129.4	5.414×10^{-3}
DPPC/	–	–	–	–	–	–	1.312×10^{-3}
RaLPS	9.00	–	–	76.5	–	–	1.128×10^{-2}
	9.00	1.70	–	61.7	23.2	–	1.818×10^{-2}
	1.01	8.93	2.33	61.0	63.1	15.8	1.944×10^{-2}

measured) does not need to be measured as long as the optimal second choice. From this, we suggest that an optimal 'model blind' measurement strategy would be to measure two contrasts (D_2O and H_2O), and either spend an equal time on both or spend slightly longer on the contrast matching the tailgroup deuteration state.

Now that we have an idea about how many contrasts to measure, we can optimize the choice of angle(s). Table 2 shows the optimized angles and counting time splits for each bilayer model, assuming a fixed time budget, using one to four angles and D_2O and H_2O contrasts. From Table 2, it appears that measuring two angles is an improvement on just one, but with any more angles, the choices are once more repeated. The results suggest that, for the DMPC and DPPC/RaLPS models, a single low angle (0.8 and 0.7°, respectively) and a single high angle (4.0°) is preferable. This is generally in line with current practices, although for OFFSPEC, angles of more than 3.0° are not commonly measured. For reference, the original experimentally measured data for the DPPC/RaLPS bilayer were measured using two angles of 0.7 and 2.3° (Clifton *et al.*, 2016). A low angle typically captures detail about the critical edge whereas a high angle provides information at high Q. What may come as a surprise is the counting times, with only about 11 and 9% of the total time budget spent measuring the low angles in each case, respectively. This could suggest that there is little information about the model parameters of interest at low Q (there is no critical edge for H_2O) and it is therefore not worthwhile measuring for very long when compared with measuring for longer at a higher angle. Alternatively, it could indicate that the information at low Q can be extracted more easily and therefore less time needs to be spent probing there. We note that if the times to change angle were included in the optimization, the results could differ slightly from those shown here. This result is also instrument dependent. For example, it is likely that a white-beam instrument with a smaller wavelength range would favour more angles, although for any instruments similar to OFFSPEC (wavelength range of [1, 14] Å) the optimal angles will be very similar. We also note that it may be beneficial to run sub-optimal angles since increased overlap of data points may add diagnostic data, *e.g.* checking that the data scale correctly with angle and that overlapping data points agree.

3.1.3. Underlayers. As detailed in Section 2.4, we were able to parameterize the bilayer models as a function of an added underlayer, whose thickness and SLD we can control. Table 3 shows the optimized underlayer properties for each bilayer model using zero to three underlayers, assuming D_2O and H_2O contrasts are being measured. As can be seen, there is a large improvement when adding an underlayer to both models. There does also seem to be an improvement when adding multiple underlayers, but the relative improvements attained are comparatively small. When considering implementation of these designs in practice, two underlayers are typically required to achieve the desired surface chemistry (Clifton *et al.*, 2015), but the additional underlayer parameters may result in the model becoming too complex for the experimentally measured data. As a result, fitting could be problematic due to large parameter uncertainties and so, for simplicity, we proceed here with the addition of a single underlayer.

We compare the values of Table 3 with those of commonly used gold and permalloy, whose SLDs are 4.7×10^{-6} and 8.4×10^{-6} Å^{-2}, respectively, and we see that these are relatively close to the optimal values for the DMPC experiments (albeit in a different order). Since most of the improvement comes from the addition of a single layer, we hypothesize that the majority of measurements made on silicon would improve if measured with a gold or permalloy underlayer, even though this is not the optimal underlayer for all circumstances. For the DMPC bilayer, we get minimum eigenvalues of 3.835×10^{-3} and 3.490×10^{-3} for the addition of single 100 Å gold and permalloy underlayers, respectively (compared with 4.694×10^{-4} with no underlayer); for the DPPC/RaLPS bilayer, we get minimum eigenvalues of 4.576×10^{-3} and 1.011×10^{-2}, respectively (compared with 1.312×10^{-3} with no underlayer). Therefore, for DMPC, it appears that either gold or permalloy will result in a significant improvement, but for DPPC/RaLPS, permalloy is preferable. To better visualize the

optimization space, Fig. 3 shows how the minimum eigenvalue changes for each SLD and thickness of the added underlayer. As can be seen, for the DMPC bilayer, the majority of optimization space is relatively flat and so most underlayer SLDs and thicknesses will result in a significant improvement. However, for the DPPC/RaLPS bilayer, this is not the case, and the choice of both conditions appears to have a more significant effect.

To illustrate the improvement achieved by underlayer addition, nested sampling was used on simulated D_2O and

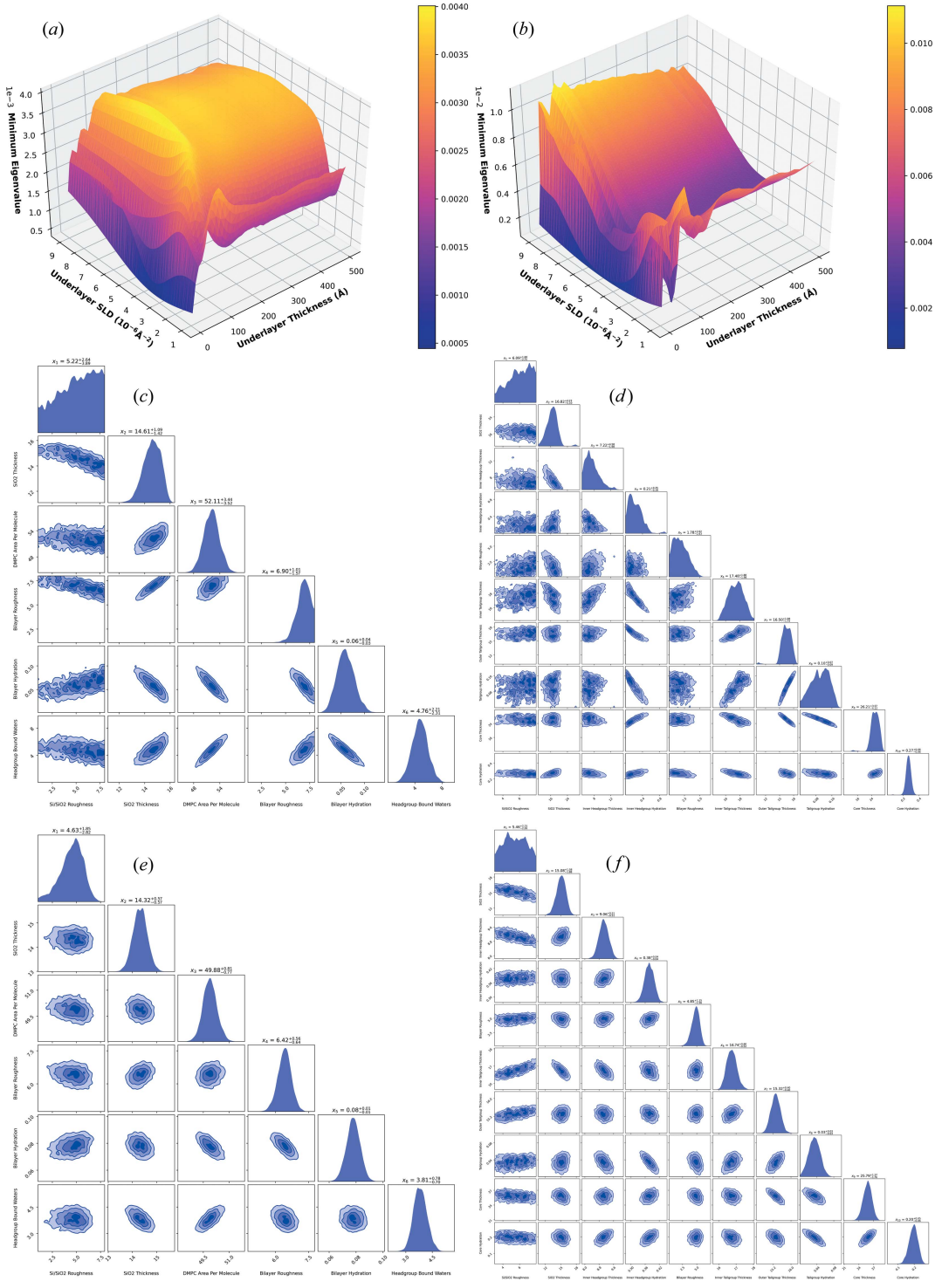


Figure 3
Plots on the left correspond to the DMPC bilayer and plots on the right to the DPPC/RaLPS bilayer. Shown [(*a*) and (*b*)] are the plots of minimum eigenvalue versus underlayer SLD and underlayer thickness for the bilayer models, assuming D_2O and H_2O are being measured. Also shown are the nested sampling corner plots from sampling simulated data of D_2O and H_2O without any underlayer [(*c*) and (*d*)] and with a single underlayer [(*e*) and (*f*)] defined by the optimized values of Table 3.

H_2O data with the default *dynesty* stopping criteria, as shown in Fig. 3. The simulation conditions were the same as previously used except only the 0.7° angle was simulated to better illustrate the improvement in parameter uncertainties and posterior distributions with the added underlayer; with two angles, the distributions are already well defined (as can be seen in Fig. 1) and so the improvement would be less noticeable. From Fig. 3, an improvement in both the estimated posterior distributions and 95%/2σ credible intervals can clearly be seen when the suggested underlayer is added.

3.2. Kinetics

Using the monolayer model of Section 2.5, the lipid APM was increased (and therefore surface excess decreased) to simulate the monolayer degrading over time. Fig. 4 shows the SLD profile for the monolayer and how the FI in the lipid APM parameter changes with measurement angle and contrast SLD. The results were obtained for two models: one with h-DPPG and the other with d-DPPG.

For both h-DPPG and d-DPPG, a contrast SLD of approximately 0×10^{-6} Å^{-2} appears to maximize the APM information. This strongly agrees with common practices where NRW (SLD of approximately 0×10^{-6} Å^{-2}) is commonly measured for its sensitivity to the water–lipid interfacial region (Clifton *et al.*, 2011). When considering angles, 1.4 and 0.4° maximize the APM information in the h-DPPG and d-DPPG models, respectively. For the d-DPPG model, there does also appear to be another solution that is only slightly worse than the optimal solution where D_2O is measured at a high angle. There is more information in the lipid APM parameter when the tailgroups are deuterated, probably due to the increased reflectivity resulting in a greater total neutron count.

3.3. Magnetism

Fig. 5 shows the experimentally fitted SLD profile and reflectivity data for the magnetic sample of Section 2.6. Also shown is how the FI in the platinum layer magnetic SLD changes with YIG and platinum layer thicknesses. The YIG thickness has virtually no effect on the FI in the platinum layer magnetic SLD. This is perhaps unsurprising, since the oscillations in reflectivity from the YIG are mostly damped at high Q where the signal from the induced magnetism would lie. However, the platinum thickness does have an effect, and maximizes the FI at around 26 Å. This is reasonably close to the fitted value of 21 Å (from the experimentally measured data) and so the original experiment appears to have fortuitously been near optimal, given the varying conditions (*i.e.* ignoring the choice of angles, counting times *etc.*). The difference in maximum and minimum FI in the Fig. 5 plot is not particularly large and so, in practice, the improvement could be diminished by other unaccounted-for factors.

Fig. 5 demonstrates the improvement obtained using the suggested platinum layer thickness by illustrating how the log ratio of likelihoods between two models, one with an induced moment of 0.01 μ_B per atom and one with no moment, changes as a function of measurement time from 1 to 100 h.

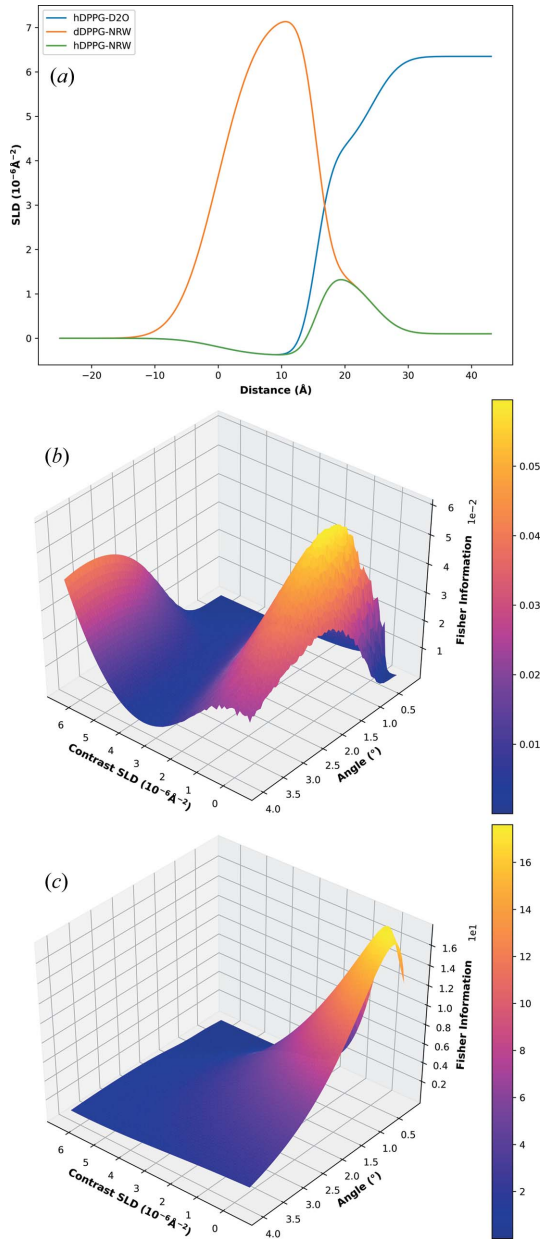


Figure 4
Shown are the model SLD profiles for the d-DPPG monolayer on NRW, h-DPPG on NRW and h-DPPG on D_2O (*a*). Also shown are the plots of FI in the lipid APM versus contrast SLD and measurement angle for the h-DPPG monolayer (*b*) and d-DPPG monolayer (*c*) models.

The plot shows two lines: one for the optimized design with 26 Å platinum layer thickness and the other for a sub-optimal design with 80 Å platinum layer thickness. As can be clearly seen, over the entire range of times under consideration, the improved design requires a lower level of counting statistics to discriminate the 0.01 μ_B per atom moment. In fact, the change of thickness away from the optimal requires that the experiment be counted for several thousand more minutes to obtain the same certainty in the result. These results also show that, with an optimized experiment, a moment of 0.01 μ_B per atom in a layer around only 20 Å thick is measurable in a reasonable time frame.

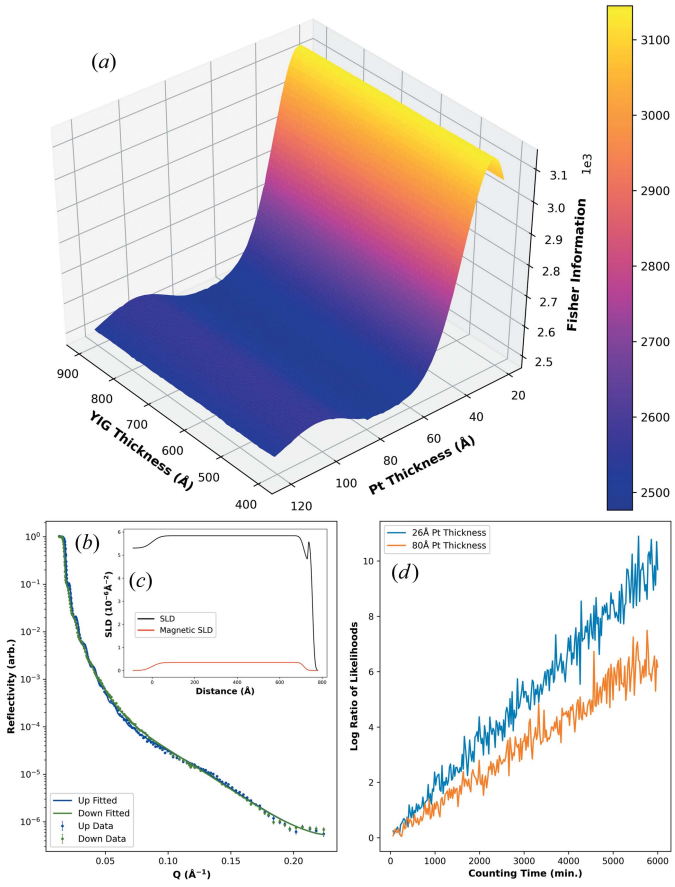


Figure 5
FI in the platinum layer magnetic SLD versus YIG and platinum layer thicknesses (*a*). Also shown are the experimentally fitted SLD and reflectivity profiles [(*b*) and (*c*), respectively]. Finally, shown is the log ratio of likelihoods between two models, one with an induced moment in the platinum layer and one with no moment, versus counting time (*d*) for two structures: one with an optimized 26 Å platinum layer thickness and the other with a sub-optimal 80 Å thickness. The counting times shown are minutes per measured spin state (*i.e.* for this model, the total times are double the values shown here).

3.4. Discussion

This work has shown that, with relatively minor tweaks to experimental procedures, experiments can be performed in shorter time frames or with greater confidence in the results of interest. In covering the wide breadth of experimental techniques presented here, we have made occasional assumptions that may mean the optimal conditions presented here are not necessarily those that you might choose to measure with, *e.g.* changing contrast has no time penalty, or measuring higher angles does not increase background significantly. We have shown that the optimal values for a large number of the variable experimental measurement conditions are highly model dependent. However, taking the example of the two bilayer models, while the optimal third contrast choice is quite different, a commonly chosen middle value of SMW avoids the complications of having to know the model in advance, with only a very minor loss of information. The shapes of many of the optimization surfaces are complex, and it may well suit the experimenter better to choose a slightly less optimal but flatter region for experimental design, for example if the behaviour is highly oscillatory. For multi-parameter problems where larger-dimensional spaces are to be probed, DE is able to provide the global optimum without exhaustive calculation; however, in these cases care should be taken to avoid sharp maxima if not all parameters can be appropriately controlled (*e.g.* film thicknesses, roughnesses, coverages or SLDs in alloys). This could be achieved either through a modification of the optimizer or, more simply, by semi-manually investigating lower-dimensional projections of the FI with individual parameters or parameter pairs in the way of Fig. 2 or Fig. 5(*a*).

We note that the increases in the minimum eigenvalue shown here correspond to better experimental outcomes; an increase by a factor of 100 in the minimum eigenvalue corresponds to a reduction in uncertainties of a factor of 10 across the worst linear combination of parameters. Additionally, since the FI increases linearly with neutron counts, an increase in the information gained is equivalent to a commensurate increase in the neutron flux. As was shown with the magnetic example, this can mean the difference between an experiment being feasible in the allotted time, or not. We should note that the maximin optimization we use here is not the only possible way of reducing the $N \times M$ matrix of the FI down to a single number (and changing this in the codebase is fairly trivial), but it does represent what we believe would be the 'expected' optimal, *i.e.* minimizing variance and covariance across all parameters of interest.

The FI framework is developed from a frequentist view, but similar Bayesian approaches have been developed, also aimed at reflectometry experiment optimization. As described in our previous work (Durant *et al.*, 2021*a*), the main difference in approaches is that the FI is calculated from the model and the neutron counts; this therefore assumes that the model describes the data but gives the benefit of never having to fit the data or sample the posteriors. The Bayesian approach instead probes the data, trying to match models to the data and reconstruct the posteriors, requiring, at the very least, fitting of the data sets and expensive calculations which would usually preclude the development of equivalent tools to those presented here. One such Bayesian approach quantifies the extractable information gain by comparing the entropies of the prior and posterior probability density functions, representing the knowledge about a sample before and after an experiment (Treece *et al.*, 2019; Heinrich *et al.*, 2020). That approach led to many of the same conclusions that are presented here about contrasts and underlayers. However, the framework requires the use of computationally expensive MCMC simulation and therefore may not be suitable for on-experiment design optimization or searching large parameter spaces.

The code for interfacing the framework, with all of the model systems presented in this work, is available in our GitHub repository (Durant *et al.*, 2021*b*). The repository also presents many of the examples discussed here as interactive *Jupyter* notebooks. These examples should be relatively easily modifiable for local applications. An incident flux file profile

for the instrument being simulated is required; however the files for several of the ISIS instruments are provided.

4. Conclusions

In this work, we demonstrated how the FI can be used to optimize experimental design across a wide range of different scientific applications in NR. We have shown that, for the two lipid bilayer systems investigated, the addition of an underlayer on the silicon substrate could provide a significant improvement, often over a factor of 10 in the minimum eigenvalue, but that the details of the optimal underlayer are model dependent. When choosing contrasts in these experiments, the first two should always be pure D_2O and H_2O; the optimal third contrast is model dependent but the commonly used SMW is a good compromise if the model is not well known. For a kinetic measurement of a monolayer on water, we showed that the optimal water contrast to measure is air matched, but that the optimal angle depends on the deuteration state of the monolayer. For magnetic measurements, we demonstrated that it is possible to measure induced moments as low as 0.01 μ_B per atom in a thin layer, as long as the sample is optimized, and that it is possible to determine how long the measurement should be conducted in order to exceed any given confidence threshold that the moment is present.

The framework is not specific to NR; any technique that can be accurately simulated, and whose error bars rely on Poisson statistics, can interface with the code. A number of *Jupyter* notebooks as well as all of the code for this work are open source and freely available on GitHub (Durant *et al.*, 2021*b*).

Acknowledgements

We thank Luke Clifton for his assistance and expertise in fitting the lipid monolayer and lipid bilayer data sets.

References

Abelès, F. (1948). *Ann. Phys.* **12**, 504–520.
Ambaye, H., Goyette, R., Parizzi, A. & Klose, F. (2008). *Neutron News*, **19**(3), 11–13.
Björck, M. & Andersson, G. (2007). *J. Appl. Cryst.* **40**, 1174–1178.
Campbell, R. A., Saaka, Y., Shao, Y., Gerelli, Y., Cubitt, R., Nazaruk, E., Matyszewska, E. & Lawrence, M. J. (2018). *J. Colloid Interface Sci.* **531**, 98–108.
Carmona Loaiza, J. M. & Raza, Z. (2021). *Mach. Learn. Sci. Technol.* **2**, 025034.
Clifton, L. A., Ciesielski, F., Skoda, M. W. A., Paracini, N., Holt, S. A. & Lakey, J. H. (2016). *Langmuir*, **32**, 3485–3494.
Clifton, L. A., Holt, S. A., Hughes, A. V., Daulton, E. L., Arunmanee, W., Heinrich, F., Khalid, S., Jefferies, D., Charlton, T. R., Webster, J. R. P., Kinane, C. J. & Lakey, J. H. (2015). *Angew. Chem. Int. Ed.* **54**, 11952–11955.
Clifton, L. A., Sanders, M. R., Hughes, A. V., Neylon, C., Frazier, R. A. & Green, R. J. (2011). *Phys. Chem. Chem. Phys.* **13**, 17153.
Cooper, J. F. K., Kinane, C. J., Langridge, S., Ali, M., Hickey, B. J., Niizeki, T., Uchida, K., Saitoh, E., Ambaye, H. & Glavic, A. (2017). *Phys. Rev. B*, **96**, 104404.
Cubitt, R., Saerbeck, T., Campbell, R. A., Barker, R. & Gutfreund, P. (2015). *J. Appl. Cryst.* **48**, 2006–2011.
Dalgliesh, R. M., Langridge, S., Plomp, J., de Haan, V. O. & van Well, A. A. (2011). *Physica B*, **406**, 2346–2349.
Doucet, M., Archibald, R. K. & Heller, W. T. (2021). *Mach. Learn. Sci. Technol.* **2**, 035001.
Duffy, L. B., Steinke, N.-J., Burn, D. M., Frisk, A., Lari, L., Kuerbanjiang, B., Lazarov, V. K., van der Laan, G., Langridge, S. & Hesjedal, T. (2019). *Phys. Rev. B*, **100**, 054402.
Durant, J. H., Wilkins, L., Butler, K. & Cooper, J. F. K. (2021*a*). *J. Appl. Cryst.* **54**, 1100–1110.
Durant, J. H., Wilkins, L. & Cooper, J. F. K. (2021*b*). *Experimental-design*, https://github.com/James-Durant/experimental-design.
Fisher, R. A. (1925). *Math. Proc. Camb. Phil. Soc.* **22**, 700–725.
Greco, A., Starostin, V., Hinderhofer, A., Gerlach, A., Skoda, M., Kowarik, S. & Schreiber, F. (2021). *Mach. Learn. Sci. Technol.* **2**, 045003.
Hastings, W. K. (1970). *Biometrika*, **57**, 97–109.
Heinrich, F., Kienzle, P. A., Hoogerheide, D. P. & Lösche, M. (2020). *J. Appl. Cryst.* **53**, 800–810.
Hughes, A. V. (2017). *RasCAL SourceForge*, https://sourceforge.net/projects/rscl/.
Inyang, O., Bouchenoire, L., Nicholson, B., Tokaç, M., Rowan-Robinson, R. M., Kinane, C. J. & Hindmarch, A. T. (2019). *Phys. Rev. B*, **100**, 174418.
Khaydukov, Y. N., Nagy, B., Kim, J.-H., Keller, T., Rühm, A., Nikitenko, Y. V., Zhernenkov, K. N., Stahn, J., Kiss, L. F., Csik, A., Bottyán, L. & Aksenov, V. L. (2013). *JETP Lett.* **98**, 107–110.
Kienzle, P. A., Maranville, B. B., O'Donovan, K. V., Ankner, J. F., Berk, N. K. & Majkrzak, C. F. (2017). *NCNR Reflectometry Software*, https://www.nist.gov/ncnr/data-reduction-analysis/reflectometry-software.
Liu, Y. & Ke, X. (2015). *J. Phys. Condens. Matter*, **27**, 373003.
Majkrzak, C. F. & Berk, N. F. (1995). *Phys. Rev. B*, **52**, 10827–10830.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.
Mironov, D., Durant, J. H., Mackenzie, R. & Cooper, J. F. K. (2021). *Mach. Learn. Sci. Technol.* **2**, 035006.
Nelson, A. R. J. & Prescott, S. W. (2019). *J. Appl. Cryst.* **52**, 193–200.
Parratt, L. G. (1954). *Phys. Rev.* **95**, 359–369.
Penfold, J., Richardson, R. M., Zarbakhsh, A., Webster, J. R. P., Bucknall, D. G., Rennie, A. R., Jones, R. A. L., Cosgrove, T., Thomas, R. K., Higgins, J. S., Fletcher, P. D. I., Dickinson, E., Roser, S. J., McLure, I. A., Hillman, A. R., Richards, R. W., Staples, E. J., Burgess, A. N., Simister, E. A. & White, J. W. (1997). *Faraday Trans.* **93**, 3899–3917.
Penfold, J., Ward, R. C. & Williams, W. G. (1987). *J. Phys. E Sci. Instrum.* **20**, 1411–1417.
Skilling, J. (2004). *AIP Conf. Proc.* pp. 395–405.
Skilling, J. (2006). *Bayesian Anal.* **1**, 833–860.
Skoda, W. A. M. (2019). *Curr. Opin. Colloid Interface Sci.* **42**, 41–54.
Speagle, J. S. (2019). *Mon. Not. R. Ast. Soc.* **493**, 3132–3158.
Storn, R. & Price, K. (1997). *J. Glob. Optim.* **11**, 341–359.
Treece, B. W., Kienzle, P. A., Hoogerheide, D. P., Majkrzak, C. F., Lösche, M. & Heinrich, F. (2019). *J. Appl. Cryst.* **52**, 47–59.
Webster, J., Holt, S. & Dalgliesh, R. (2006). *Physica B*, **385–386**, 1164–1166.
Welbourn, R. J. L. & Clarke, S. M. (2019). *Curr. Opin. Colloid Interface Sci.* **42**, 87–98.
Zhan, X. Z., Li, G., Cai, J. W., Zhu, T., Cooper, J. F. K., Kinane, C. J. & Langridge, S. (2019). *Sci. Rep.* **9**, 6708.
Zhang, Y., Parnell, A. J., Pontecchiani, F., Cooper, J. F. K., Thompson, R. L., Jones, R. A. L., King, S. M., Lidzey, D. G. & Bernardo, G. (2017). *Sci. Rep.* **7**, 44269.