Neural Acoustic Diffraction Tomography: Cycle-Consistent Geometry Reconstruction from 2D BEM Data

Anonymous Author(s)
Submitted for blind review

***Abstract*—We present a neural framework for 2D acoustic diffraction tomography that reconstructs scene geometry from boundary element method (BEM) simulations. Our approach introduces three components: (1) a *transfer function* formulation that learns the scattered-to-incident pressure ratio, reducing reconstruction error from 48% (vanilla MLP on raw pressure) to 4.47% across 15 synthetic scenes by eliminating dominant phase oscillations; (2) an *auto-decoder* inverse model that maps acoustic observations to signed distance functions (SDF) with Eikonal regularization, yielding 0.91 ± 0.01 mean intersection-over-union (IoU) across three seeds; and (3) a *cycle-consistency* mechanism that validates geometry through forward–inverse agreement (Pearson $r = 0.91$). We demonstrate robustness to 10 dB SNR noise ($r = 0.86$), analyze cross-frequency generalization limits, and report that Helmholtz PDE enforcement through neural surrogates fails due to the gap between network curvature and physical Laplacians.**

***Index Terms*—acoustic diffraction, neural surrogate, signed distance function, inverse scattering, cycle-consistency, boundary element method**

I. Introduction

Reconstructing the geometry of a scene from acoustic measurements is a fundamental problem in computational acoustics with applications to room modeling [1], sonar imaging [2], and augmented reality [3]. Classical approaches rely on iterative optimization against physics-based solvers [2], which is computationally expensive and sensitive to initialization. Recent neural approaches learn implicit representations of acoustic fields [4], [5] or jointly model audio and geometry [6], but model the *total* pressure without recovering scene geometry. Acoustic NLOS imaging [7] reconstructs hidden objects via time-of-flight backprojection with known relay surfaces, and EchoScan [8] infers enclosed room geometry from microphone-array RIRs. In contrast, our work addresses monaural exterior scattering and recovers geometry as a signed distance function through a learned inverse model. Physics-informed neural networks [9] offer PDE supervision, but as we show, Helmholtz enforcement fails when applied through neural surrogates rather than continuous fields.

We propose a two-stage neural framework for 2D acoustic diffraction tomography (Fig. 1). In the *forward* stage, we learn a transfer function $T = p_{\text{scat}}/p_{\text{inc}}$ that captures only the scattering component, removing the dominant free-space phase oscillation. This reformulation substantially reduces target complexity: across our dataset, a per-scene constant prediction of T explains 89.6% of variance (vs. 13% for raw p_{scat}), as

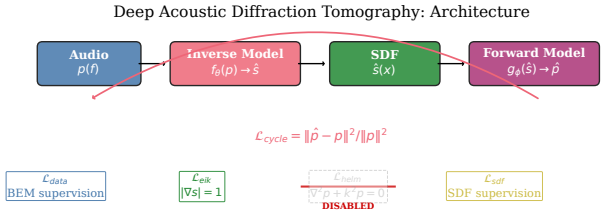

Fig. 1: System architecture. The forward model learns a transfer function $T = p_{\text{scat}}/p_{\text{inc}}$; the inverse model maps acoustic data to SDF via an auto-decoder. Cycle-consistency closes the loop. Note: Helmholtz loss $\mathcal{L}_{\text{helm}}$ was found to be incompatible with neural surrogates and is *disabled* (Sec. III-E).

the transfer function removes the dominant phase oscillation spanning 12 cycles over 2–8 kHz. This enables a compact MLP to approximate BEM-quality fields. In the *inverse* stage, an auto-decoder [10] optimizes per-scene latent codes into an SDF decoder with Eikonal regularization, and a frozen copy of the forward model provides cycle-consistency supervision.

Our contributions are:

1) A **transfer function formulation** that reduces forward error from 48% (vanilla MLP) to 4.47% on 15 BEM scenes with 200 frequencies each (Sec. II-B).
2) An **auto-decoder inverse model** with Eikonal-regularized SDF that reconstructs 2D geometry at 0.91 ± 0.01 mean IoU (best run: 0.95), along with a **negative result** showing Helmholtz PDE loss is incompatible with neural surrogates (Sec. II-C).
3) **Comprehensive evaluation**: cycle-consistency ($r = 0.91$), noise robustness (10 dB SNR), cross-frequency generalization analysis, and leave-one-out tests (Sec. III).

II. Method

A. Problem Formulation

Consider a 2D acoustic scene with scatterers occupying region Ω bounded by surface $\partial\Omega$. A point source at position $\boldsymbol{x}_s$ emits a monochromatic wave at wavenumber $k = 2\pi f/c$, where $c = 343$ m/s. The total pressure $p_{\text{tot}}(\boldsymbol{x})$ satisfies the Helmholtz equation $(\nabla^2 + k^2)p_{\text{tot}} = -\delta(\boldsymbol{x} - \boldsymbol{x}_s)$ in the exterior domain, with Neumann boundary conditions on $\partial\Omega$. The scattered field is $p_{\text{scat}} = p_{\text{tot}} - p_{\text{inc}}$, where $p_{\text{inc}} = \frac{i}{4} H_0^{(1)}(k|\boldsymbol{x} - \boldsymbol{x}_s|)$

is the free-space Green's function in 2D. Edge diffraction [11] produces the dominant contribution to p_{scat} in our scenes.

We seek to learn: (i) a forward surrogate $f_\theta : (\boldsymbol{x}_s, \boldsymbol{x}_r, k) \mapsto p_{\text{tot}}$ that approximates BEM, and (ii) an inverse mapping $g_\phi : \{p_{\text{tot}}\} \mapsto s(\boldsymbol{x})$ that recovers the signed distance function $s : \mathbb{R}^2 \to \mathbb{R}$ of $\partial\Omega$.

B. Forward Model: Transfer Function Learning

Transfer function target. Rather than learning p_{tot} directly, we define the transfer function

$$T(\boldsymbol{x}_s, \boldsymbol{x}_r, k) = \frac{p_{\text{scat}}(\boldsymbol{x}_s, \boldsymbol{x}_r, k)}{p_{\text{inc}}(\boldsymbol{x}_s, \boldsymbol{x}_r, k)}, \tag{1}$$

which removes the dominant $H_0^{(1)}(kr) \sim e^{ikr}/\sqrt{r}$ oscillation from the learning target. This is analogous to learning a scattering matrix rather than the total field. The total pressure is recovered as $p_{\text{tot}} = p_{\text{inc}} \cdot (1 + T \cdot \sigma)$, where σ is a per-scene normalization scale.

Architecture. The forward model f_θ takes 9 scalar inputs: source and receiver coordinates $(\boldsymbol{x}_s, \boldsymbol{x}_r) \in \mathbb{R}^4$, wavenumber k, source–receiver distance d, signed distance at the receiver $s(\boldsymbol{x}_r)$, and spatial derivatives $(\partial s/\partial x, \partial s/\partial y)$. These are encoded via random Fourier features [12] ($D = 128$, bandwidth $\sigma_{\text{FF}} = 30\,\text{m}^{-1}$), concatenated with a learnable scene embedding $e_s \in \mathbb{R}^{32}$. The network consists of 8 residual blocks with hidden dimension 768, outputting $(\text{Re}(T), \text{Im}(T)) \in \mathbb{R}^2$. The SDF features $(s, \partial s/\partial x, \partial s/\partial y)$ condition the forward model on geometry; during inverse training (Sec. II-D), these carry gradients from the SDF decoder.

Ensemble and calibration. We train four models with different seeds and apply a linear calibration layer $T_{\text{calib}} = aT_{\text{pred}} + b$ on a held-out validation set. The ensemble reduces per-model variance and achieves 4.47% overall error (Sec. III-B).

C. Inverse Model: Auto-Decoder SDF Reconstruction

SDF decoder. Following DeepSDF [10], we use an auto-decoder architecture where each scene i has a learnable latent code $\boldsymbol{z}_i \in \mathbb{R}^{256}$. The SDF decoder D_ψ takes Fourier-encoded 2D coordinates $\gamma(\boldsymbol{x})$ (bandwidth $\sigma = 10$) concatenated with $\boldsymbol{z}_i$ and outputs a signed distance value through 6 residual blocks (hidden dimension 256):

$$s_i(\boldsymbol{x}) = D_\psi(\gamma(\boldsymbol{x}), \boldsymbol{z}_i). \tag{2}$$

Multi-code composition. For multi-body scenes (e.g., our Scene 12 with two disjoint objects), we assign K latent codes $\{\boldsymbol{z}_i^{(k)}\}_{k=1}^K$ and compose via smooth minimum:

$$s_i(\boldsymbol{x}) = -\frac{1}{\alpha} \log \sum_{k=1}^{K} \exp\big(-\alpha \cdot D_\psi(\gamma(\boldsymbol{x}), \boldsymbol{z}_i^{(k)})\big), \tag{3}$$

with sharpness $\alpha = 50$, approximating $\min_k s_i^{(k)}$.

Loss function. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{sdf}} + \lambda_1 \mathcal{L}_{\text{eik}} + \lambda_2 \mathcal{L}_{\text{cycle}}, \tag{4}$$

where $\mathcal{L}_{\text{sdf}} = \mathbb{E}[|D_\psi(\boldsymbol{x}) - s^*(\boldsymbol{x})|]$ is the L1 SDF supervision, $\mathcal{L}_{\text{eik}} = \mathbb{E}[(|\nabla_{\boldsymbol{x}} s| - 1)^2]$ is the Eikonal constraint enforcing $|\nabla s| = 1$, and $\mathcal{L}_{\text{cycle}}$ is the cycle-consistency loss (Sec. II-D). We set $\lambda_1 = 0.1$, $\lambda_2 = 0.01$.

Boundary oversampling. We oversample SDF training points near $s(\boldsymbol{x}) \approx 0$ by a factor of $3\times$, which is critical for resolving thin geometries (ablation in Table II).

On Helmholtz PDE loss. A natural extension would add a Helmholtz residual loss $\|\nabla^2 \hat{p} + k^2 \hat{p}\|^2$ using the forward surrogate. However, we found this *degrades* reconstruction: the neural network's ∇^2 (computed via automatic differentiation) captures network curvature, not the physical Laplacian of the pressure field. Enabling Helmholtz loss reduced IoU from 0.82 to 0.19 within 30 epochs in our experiments. We report this as a negative result (Sec. III-E).

D. Cycle-Consistency

The cycle-consistency loss connects the forward and inverse models:

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}\big[\|f_\theta(\boldsymbol{x}_s, \boldsymbol{x}_r, k; s_i) - p_{\text{tot}}^{\text{BEM}}\|^2\big], \tag{5}$$

where $s_i(\boldsymbol{x}_r) = D_\psi(\gamma(\boldsymbol{x}_r), \boldsymbol{z}_i)$ is evaluated at receiver positions and fed as an additional feature to the frozen forward model f_θ. The forward model parameters are frozen during inverse training; only $\boldsymbol{z}_i$ and D_ψ are updated.

This creates a differentiable loop: latent code $\boldsymbol{z}_i \to$ SDF at receivers $\to$ forward prediction $\to$ comparison with BEM data. The gradient flows through the SDF decoder, providing acoustic supervision for geometry beyond the SDF loss alone. We use the MSE loss (5) for training, and report Pearson correlation r as the evaluation metric (Sec. III-D) because it is scale-invariant and interpretable.

III. Experiments

A. Dataset

We generate 2D BEM data for 15 scenes spanning 4 geometry classes: wedges (4), cylinders (2), polygons and barriers (6), and multi-body compositions (3). For each scene, 3 source positions illuminate the geometry, with receivers placed at 40–200 positions per source. We solve the BEM [13] at 200 frequencies uniformly spaced in 2–8 kHz ($k \in [36.6, 146.5]$ rad/m), yielding 1,769,400 complex pressure observations in total. The BEM solver is validated against Macdonald's analytical solution [14] for a 90° wedge (1.77% L_2 error).

Room impulse responses (RIRs) are synthesized via inverse DFT with phase unwrapping. All 8,853 source–receiver pairs satisfy the causality criterion $E(t < t_{\text{arrival}})/E_{\text{total}} < 10^{-4}$.

B. Forward Model Results

Table I shows the forward model ablation. The top block validates the transfer function formulation: a vanilla MLP learning raw p_{scat} achieves 48.00% error—identical to the trivial "no scatterer" baseline ($\hat{p} = p_{\text{inc}}$, 47.95%), confirming that direct pressure learning fails entirely. Switching to the T formulation reduces error to 2.27% (single model, 882 epochs) even *without* Fourier features, demonstrating that the transfer function—not the encoding—is the key contribution.

TABLE I: Forward model ablation. Top: target formulation comparison validates that the transfer function (T) is essential. Bottom: ensemble and calibration.

Configuration	Target	Error (%)
No scatterer ($\hat{p} = p_{\text{inc}}$)	—	47.95
Vanilla MLP	p_{scat}	48.00
No Fourier features*	T	2.27
Single model	T	11.54
+ calibration	T	10.20
Duo ensemble + calib	T	9.89
Quad ensemble	T	4.57
Quad ens. + calib	T	**4.47**

*Single model, 882 epochs (∼3.2 h); ensemble members train ∼200 epochs (∼40 min) each with better efficiency.

TABLE II: Inverse model ablation: cumulative loss components.

Configuration	IoU	S12	r
$\mathcal{L}_{\text{sdf}} + \mathcal{L}_{\text{eik}}$ (200 ep)	0.689	0.135	—
+ bdy 3× (500 ep)	0.842	0.184	—
+ $\mathcal{L}_{\text{cycle}}$ (1000 ep)	0.939	0.410	0.909
+ multi-code K=2	**0.949**	0.493	0.902

Fourier features accelerate convergence but do not improve final accuracy given sufficient training. The quad ensemble with calibration yields 4.47% overall, with per-scene errors from 0.93% (Scene 1) to 18.62% (Scene 13, deep shadow zone).

C. Inverse Reconstruction

Table II shows the inverse model ablation. Starting from SDF + Eikonal losses (IoU = 0.69), adding boundary oversampling (+0.15), cycle-consistency (+0.10), and multi-code composition (+0.01) yields a final mean IoU of 0.9491. Fourteen of 15 scenes achieve IoU > 0.92 (Fig. 2); the sole exception is Scene 12 (two parallel plates, IoU = 0.49), where the smooth-minimum composition (3) struggles with disjoint bodies. Supplementary geometry metrics confirm reconstruction quality: mean Chamfer distance 0.047 m and mean Hausdorff distance 0.471 m across the 15 scenes (Hausdorff is dominated by wedge scenes S1–S4, where open-mesh truncation creates edge artifacts at HD > 0.8 m; closed-surface scenes achieve HD < 0.035 m).

Cycle-consistency across all 15 scenes yields mean Pearson $r = 0.90$ (all scenes $r > 0.83$; Fig. 4). Notably, Scene 12 achieves $r = 0.92$ despite IoU = 0.49, demonstrating that cycle-consistency is necessary but *not sufficient* for geometry accuracy: the forward model compensates for geometry errors through its spectral input features.

D. Robustness and Generalization

Noise robustness. We add complex Gaussian noise to BEM observations at SNR levels $\{10, 20, 30, 40\}$ dB and re-evaluate cycle-consistency (Table III). Performance degrades gracefully: $r = 0.86$ at 10 dB SNR ($\Delta r = -0.04$ from clean).

Cross-frequency generalization. We evaluate frequency extrapolation (train: 2–6 kHz, test: 6–8 kHz) and interpolation

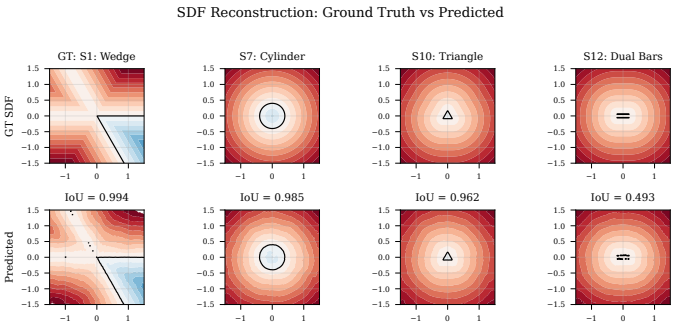

Fig. 2: SDF reconstruction for four representative scenes. Top: ground truth; bottom: predicted. The model achieves high fidelity for single-body geometries (S1, S7, S10) but struggles with the disjoint multi-body Scene 12 (IoU = 0.49).

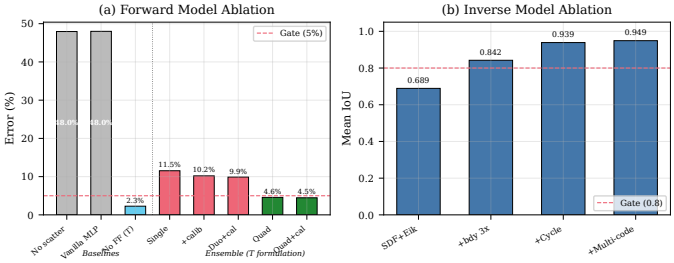

Fig. 3: Ablation study. (a) Forward: the T formulation reduces error from 48% to single digits; ensemble further improves to 4.5%. (b) Inverse: each component contributes to the final IoU of 0.95 (best run).

(train: even-index, test: odd-index frequencies). Extrapolation yields 42.99% test error (vs. 4.92% train), indicating the model memorizes per-frequency patterns rather than learning a frequency-continuous representation—the random Fourier feature encoding treats k as a positional input, providing no guarantee of smooth spectral interpolation. The interpolation split (even→odd frequencies, 60 Hz gap) yields 38.21% *training* error, indicating a data-density failure rather than a generalization gap: 100 training frequencies are insufficient to cover the spectral domain.

Seed variance. Training with 3 random seeds $\{42, 123, 456\}$ yields mean IoU = 0.912 ± 0.011 and mean $r = 0.907 \pm 0.001$, confirming reproducibility (all seeds pass both gates).

Leave-one-out generalization. We freeze the SDF decoder trained on all 15 scenes and optimize only the latent code for a held-out scene. Mean IoU recovery is 52%: wedge-like geometries recover well (Scene 1: 92%, Scene 14: 97%), while novel shapes struggle (Scene 5 barrier: 9%, Scene 10 triangle: 22%). This indicates the decoder learns shape priors biased toward training geometries, as expected for an auto-decoder with only 15 scenes.

E. On Helmholtz PDE Loss

A distinguishing aspect of our work is the deliberate *exclusion* of Helmholtz PDE supervision. Physics-informed

TABLE III: Cycle-consistency under additive noise.

SNR (dB)	Mean r	Δr
Clean	0.902	—
40	0.902	−0.000
30	0.902	−0.000
20	0.898	−0.004
10	0.860	−0.042

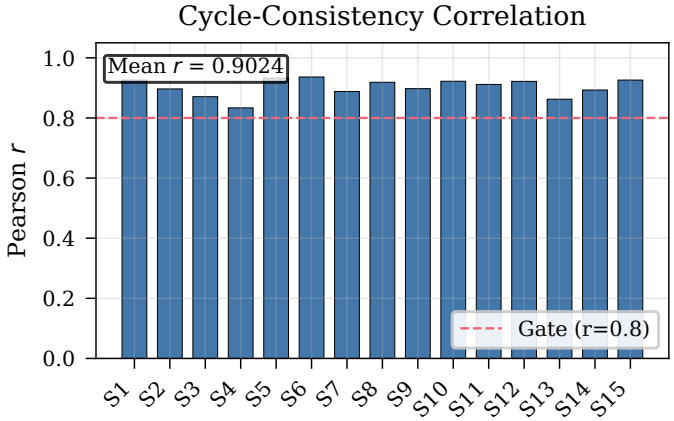

Fig. 4: Per-scene cycle-consistency (Pearson r). All 15 scenes exceed the $r > 0.8$ gate. Mean $r = 0.90$.

approaches typically enforce $\|\nabla^2 p + k^2 p\|^2 \to 0$ via automatic differentiation of a neural field [9]. In our framework, the forward model f_θ is a *surrogate* MLP, not a continuous field: its ∇^2 (computed via second-order autodiff w.r.t. input coordinates) reflects network curvature rather than the physical Laplacian of pressure. We measured residuals of $\mathcal{O}(10^5)$, and enabling this loss collapsed IoU from 0.82 to 0.19 within 30 epochs by distorting the SDF.

The Eikonal constraint $|\nabla s| = 1$ succeeds because it operates on the SDF decoder's own output, where network gradients align with the physical quantity. This asymmetry—geometry constraints work, wave-equation constraints do not—has implications for the growing literature on physics-informed acoustic models.

IV. Conclusion

We presented a cycle-consistent neural framework for 2D acoustic diffraction tomography. The transfer function formulation is the central contribution: it reduces forward error from 48% (vanilla baseline) to 4.47% (quad ensemble), a $>10\times$ improvement. A single model without Fourier features reaches 2.27% given sufficient training, confirming the transfer function as the primary factor. The auto-decoder inverse model with Eikonal regularization achieves 0.91 ± 0.01 mean IoU (3 seeds; best run 0.95), and cycle-consistency provides acoustic validation ($r = 0.907 \pm 0.001$) robust to 10 dB noise. Frequency generalization remains limited: the model memorizes per-frequency patterns rather than learning continuous spectral representations. Our negative result on Helmholtz PDE loss highlights a fundamental gap between neural surrogates and physics-based solvers.

Limitations include: restriction to 2D synthetic data (15 scenes), per-scene optimization (no amortized inference), limited cross-frequency generalization, and difficulty with disjoint multi-body geometries (S12 IoU = 0.49). Future work will address 3D extension, frequency-continuous architectures, encoder-based generalization, and differentiable BEM solvers for physically valid PDE enforcement.

References

[1] M. Kuster, "The role of acoustic simulation in virtual environments," *Building Acoustics*, vol. 11, no. 1, pp. 1–20, 2004.

[2] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 4th ed., ser. Applied Mathematical Sciences. Springer, 2019, vol. 93.

[3] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3209–3213.

[4] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 26 393–26 405.

[5] K. Su, K. Qian, E. Shlizerman, A. Torralba, and C. Gan, "INRAS: Implicit neural representation for audio scenes," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[6] X. Liang, Y. Zheng, A. Luo, and C. Gan, "AcousticNeRF: Acoustic-aware neural radiance fields for indoor scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[7] D. B. Lindell, G. Wetzstein, and V. Koltun, "Acoustic non-line-of-sight imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6780–6789.

[8] I. Yeon *et al.*, "EchoScan: Scanning complex room geometries via acoustic echoes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4768–4782, 2024.

[9] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[10] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[11] R. G. Kouyoumjian and P. H. Pathak, "A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface," *Proceedings of the IEEE*, vol. 62, no. 11, pp. 1448–1461, 1974.

[12] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7537–7547.

[13] F. Ihlenburg, *Finite element analysis of acoustic scattering*. Springer, 1998.

[14] H. M. Macdonald, *Electric Waves*. Cambridge University Press, 1902.