

A Simple, Modular Pipeline for Automated EIS Equivalent-Circuit Fitting via PCA- Space Nearest-Neighbour Matching, Validated Against Manual Fitting

Sanja Martinez

`sanja.martinez@fkit.unizg.hr`

University of Zagreb

Lea Hrženjak

University of Zagreb

Article

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11118325/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Simple, Modular Pipeline for Automated EIS Equivalent-Circuit Fitting via PCA-Space Nearest-Neighbour Matching, Validated Against Manual Fitting

Sanja Martinez^{*}, Lea Hrženjak

University of Zagreb, Faculty of Chemical Engineering and Technology, Marulićev trg 19,
10000 Zagreb, Croatia

^{*}Corresponding author: Sanja Martinez (sanja.martinez@fkit.unizg.hr)

Abstract

Automated, data-driven equivalent-circuit fitting of electrochemical impedance spectroscopy (EIS) data requires deciding which circuit topology best describes a spectrum and what starting parameters to use. Because distinct topologies can produce near-identical or mathematically identical spectra, committing to a single topology before evaluating fit quality risks discarding the valid answer. Here we present a pipeline that avoids this early commitment, tolerating substantial overlap between similar topologies without loss of accuracy. A synthetic reference database, generated analytically from a small number of manually fitted seed spectra, is projected into a shared principal component analysis (PCA) space together with any spectrum to be fitted. A coarse nearest-neighbour search shortlists candidates and supplies each with physically plausible starting parameters, which are then independently refined by complex nonlinear least squares (CNLS), with the lowest-chi-squared topology retained and competing candidates kept visible. Validated against independent manual fits across reference, known, unknown, and genuinely unknown measured spectra, the automated fits agreed with the manual fits within 1% for most parameters and achieved lower chi-squared and parameter errors throughout, taking 5-6 seconds per spectrum. A coverage gap for out-of-envelope spectra was resolved by adding one seed spectrum, leaving all prior fits unchanged and demonstrating modular extensibility.

Introduction

Automated interpretation of electrochemical impedance spectroscopy (EIS) data has recently become an active area of research, with a growing body of work seeking to automate the equivalent-circuit fitting process traditionally performed by hand. Automating EIS fitting workflow requires solving two related but distinct sub-problems: identifying which equivalent-circuit topology best describes the spectrum and supplying a physically plausible set of starting parameter values for whichever nonlinear fitting routine is used to refine that topology's components. Paradoxically, the main challenge, non-uniqueness of the fitting result, is not inherent to automation, but to the fitting itself. This non-uniqueness takes several forms. Distinct topologies can

produce mathematically identical impedance spectra [1], two or more time constants within the same spectrum can lie close enough together that their impedance arcs merge into a single arc and can no longer be reliably separated [2], a single circuit element's parameters can trade off against each other, or against an adjacent element, when the measured frequency range does not fully excite the associated relaxation [3], and a parameter's characteristic frequency can fall entirely outside the measured range, leaving it effectively unconstrained by the data regardless of the circuit topology chosen [4].

Nearly all automated fitting methods commit to a single predicted topology before, or independently of, evaluating how well that topology actually fits the measured spectrum. One recent exception retains a full probability distribution across candidate topologies rather than collapsing to a single one, intended to narrow rather than dictate the downstream choice of circuit [5]. Supervised classifiers such as support vector machines [6], k-nearest-neighbours [7], artificial neural networks [8,9], and random forests [4] predict a single equivalent circuit from spectral features, but handle the parameter-initialization sub-problem very differently from one another. Some train a second, separate regression model to predict starting parameters [8,9]. Some query the same model twice for the two tasks, as in the large-language-model workflow of Bongiorno et al. [11]. Others predict topology alone and hand the spectrum to a downstream commercial fitting package (ZSimpWin) without reporting how that package is itself initialized [4]. Reinforcement-learning approaches that construct circuits sequentially without any pre-training corpus, such as Auto-EIS [10] and AutoREC [12], fit parameters at every step of circuit construction. The Complex Nonlinear Least Squares (CNLS) [13] initialization used for such per-step fits, and in the ground-truth labelling step of Sáez-Pardo et al. [5], is similarly left unreported. Robust initialization of the CNLS solver has itself been the subject of dedicated hybrid algorithmic approaches [14]. One recent sequence-generation approach omits parameter estimation from the pipeline entirely, returning only a candidate topology string and explicitly deferring numerical fitting to future work [15].

Beyond this shared non-uniqueness problem, automation introduces a challenge of its own: performance is tightly coupled to the size and representativeness of the data the method has seen. Classifiers trained purely on synthetic spectra reach high accuracy on synthetic test sets [8] (about 95% on well-separated topologies) but are explicitly left untested on real, noisy data. When trained instead on a modest set of real spectra from a single instrument, classification accuracy has been shown to collapse from 99.17% to 46.59% when the same network is applied to spectra from a different potentiostat [9], a failure caused by the network memorising instrument-specific noise rather than the underlying physics.

Reinforcement-learning approaches reduce the labelled-data burden, since Auto-EIS and AutoREC need no pre-training corpus at all, but the cost shows up elsewhere. AutoREC's own dimensionality-reduction analysis shows real experimental spectra lying near the boundaries of, or between, the clusters formed by its synthetic training topologies, and training a single circuit-generation policy takes on the order of 18 CPU-hours [12]. Large language model approaches remove the need for training data entirely but introduce a substantial cost per spectrum instead (about 2.17 s and non-trivial API charges per call), along with measurable run-to-run instability. Three-way prompt agreement was as low as 0.76 for less explicit prompt designs [11].

The challenges in autofitting equivalent circuits to EIS data have recently been summarized by Orazem and Ulgut [1], who propose guidelines for developing AI algorithms for impedance analysis. In their view, such a tool should not try to output a single best model but should instead show the full set of mathematically equivalent circuits consistent with the data, suggest possible physical interpretations for each, noting that one model may itself have more than one physical explanation, and propose further experiments, for example at different potentials, temperatures, or electrode dimensions, that could help discriminate between the surviving candidates. Such a tool should guide the electrochemist rather than replace their judgement, since a single impedance spectrum, having only frequency as an independent variable, cannot on its own give the chemical specificity needed for a unique mechanistic assignment.

To demonstrate an approach built around these guidelines, this study uses a synthetic reference database generated analytically from a small number of manually fitted seed measurements, with topologies following exactly from the generating equation. A coarse, computationally inexpensive nearest-neighbour search in a reduced Principal Component Analysis (PCA) [16] representation of this database jointly addresses both sub-problems from a single retrieval step. Because every database entry carries both a known topology label and the exact parameter values used to generate it, the retrieved neighbours shortlist candidate topologies and supply physically plausible starting parameters for each at the same time. Every topology represented among the retrieved neighbours is then independently refined by CNLS from several such candidate starting points, and the final assignment is made afterward, based on fit quality (reduced χ^2). The resulting output retains all competing topologies that survive the refinement step, together with their fit quality, allowing a human expert to make the final selection when several candidates remain closely matched. This separates the cheap, candidate-generation step from the expensive, physically grounded decision step, so that a misleading nearest-neighbour match costs, at worst, a few wasted CNLS attempts run in parallel. This property, as we show, allows the pipeline to tolerate substantial topology overlap in the underlying PCA representation without measurable cost to the final fit, and offers a practical way of living with the non-uniqueness described above, leaving the final decision to complementary physical measurements and/or human

judgement. Representative examples of this output format, showing all surviving candidate topologies side by side with their reduced χ^2 and parameter tables, are provided for every fitted spectrum in this work in the accompanying data repository (see Data Availability).

Existing autofitting solutions also differ widely in how general they are. Some are designed to construct or select from an open-ended space of circuit topologies, in principle covering any equivalent circuit. Others are restricted from the outset to a small, fixed set of candidate topologies chosen for a specific system, such as two circuits representing intact versus defective coatings [9], or five circuits each tied to a particular corrosion mechanism [4]. The approach presented here falls closer to the latter group in its present form but is modular by design: the reference database can be easily extended with additional seed measurements and their synthetic neighbourhoods to cover new topologies or systems, without requiring any change to the existing database or matching pipeline. Rather than reporting only overall success rates, we explicitly characterise the conditions under which the pipeline underperforms. In each case, we show that the failure comes from a specific, locally correctable property, not affecting the pipeline itself.

Workflow and data for automatic fitting

Validation of the PCA-space distance-matching approach involved applying three levels of testing to a single underlying pipeline, implemented as a workflow in Orange [17] (Figure 1).

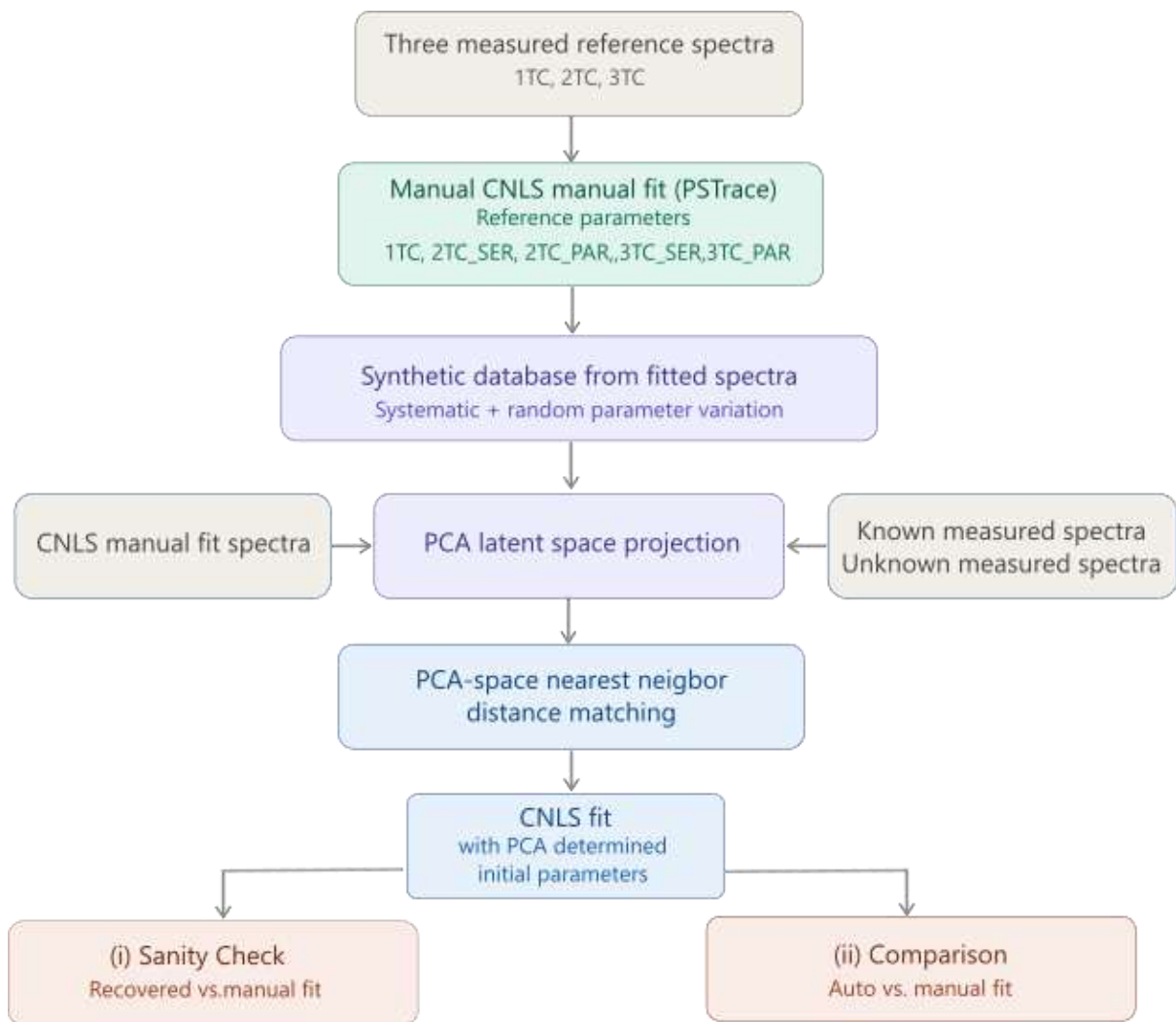


Figure 1. Workflow for constructing the synthetic reference database and validating PCA-space nearest-neighbour (NN) lookup as the strategy for topology narrowing and initialization for automated CNLS-based fitting, via both a sanity check against the recovered reference fits and a comparison against independent manual fits of known and unknown measured spectra.

The synthetic reference database was constructed from three measured reference spectra (1TC, 2TC, 3TC), each manually fitted via CNLS in PSTrace. Because the series and nested/parallel topologies cannot be distinguished a priori from spectral shape alone, the 2TC and 3TC spectra were each independently fitted under two topological assumptions, yielding five reference parameter sets in total (1TC, 2TC_SER, 2TC_PAR, 3TC_SER, 3TC_PAR). Here, TC denotes the number of time constants and SER/PAR the series or nested/parallel topology.

The database is projected into PCA space once (Figure 2). Any spectrum to be fitted is then processed through a parallel branch applying the same transformations: per-block normalization [18] using the same normalization parameters fitted on the reference database, and projection into the existing PCA space via Apply Domain Orange widget. Both branches enter the nearest neighbour distance matching of the spectrum to be fitted against the synthetic database. Topology narrowing arises from which topologies are represented among the 400 nearest neighbours, and for each topology passed on, CNLS is performed, initialized by the corresponding matched neighbour's parameters. This allows an arbitrary number of spectra to be fitted independently through this same, unchanged path. For each retained topology, CNLS refinement was performed independently from the five nearest database neighbours, and the attempt achieving the lowest reduced χ^2 was pronounced a winner and retained along with the results for other topologies. This extends the candidate-suite, reducing sensitivity to an individual unlucky nearest-neighbour match while still allowing every surviving topology to compete post hoc on an equal footing.

The three validation levels (Figure 1) differ only in the nature of the input spectrum and in what the refined result is compared against: (i) a sanity check, in which each of the manually fitted reference spectra are passed through the pipeline as an input, and the CNLS recovered parameters are compared against the original manual fits, confirming correct recovery within the synthetic database, (ii) known measured spectra, the actual measurements underlying the manually fitted reference spectra are each passed through the pipeline as an input, and the CNLS parameters are compared against the same manual fit, and (iii) unknown measured spectra are passed through the fitting branch of the pipeline as an input, and the same comparison against their independent manual fits is used to assess generalization beyond the reference measurements themselves.

In this framework, PCA-space distance serves as a computationally cheap, ad hoc coarse filter. Instead of attempting to identify the correct topology directly, it is used to shortlist plausible candidates and to supply each surviving topology with a physically reasonable set of initial parameters for CNLS refinement. This is what makes it feasible to independently fit the full impedance spectrum for every retained topology, an approach that would be considerably more expensive if the same optimization had to be run from arbitrary or randomly chosen starting points. Topology selection itself is then made post hoc, based on which candidate best matches the input spectrum (lowest χ^2), not on which candidate happened to be closest in the reduced PCA representation.

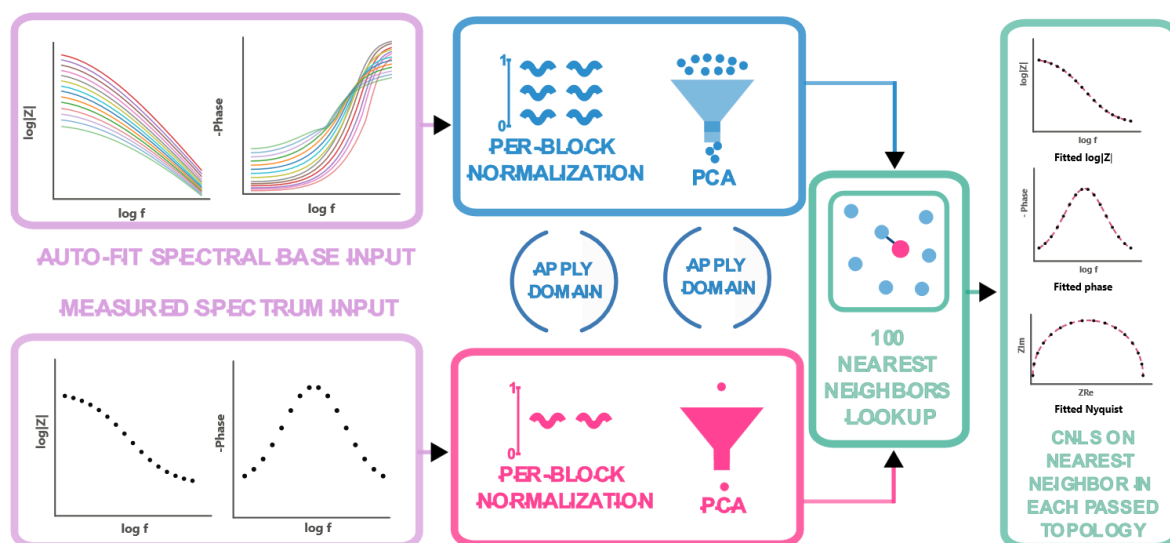


Figure 2. Detailed schematic of the PCA-space nearest-neighbour matching and automated fitting pipeline. Both the synthetic reference database (auto-fit spectral base input) and any input test spectrum (measured spectrum input) undergo identical per-block normalization and PCA projection. The test spectrum is mapped into the existing PCA space via Apply Domain, ensuring both are represented in a common latent space. The 400 nearest database neighbours to the test spectrum are then retrieved, and CNLS fitting is performed, initialized by the retrieved five nearest matched neighbours for every equivalent-circuit topology represented among them. The topology giving the lowest reduced χ^2 is selected as the final fit, shown on the right as the fitted $\log|Z|$, phase, and Nyquist spectra.

The EIS spectra used throughout this work were recorded on a PalmSenS4 potentiostat/FRA instrument. The specific materials and systems measured are not detailed here, as they are not relevant to the validation of the pipeline itself. Frequencies were sampled at five points per decade over the range 10^5 to 10^{-2} Hz, matching the fixed frequency grid used throughout the reference database and matching pipeline. Spectra truncated from either end of this range or recorded over a narrower frequency window, can still be used as input, provided the frequencies present coincide exactly with points on this fixed grid and form an unbroken subrange with no missing points in between.

Synthetic database formation

Three structural classes of equivalent circuits were considered, differing in the number of TCs needed to describe the impedance response: one (1TC), two (2TC), and three

(3TC). Each class reflects a different number of distinguishable relaxation processes contributing to the impedance spectrum (e.g., coating capacitance/pore resistance, charge-transfer/double-layer processes, or diffusion-related contributions), each represented by a parallel combination of a constant-phase element (Q_n) and a resistance (R_n).

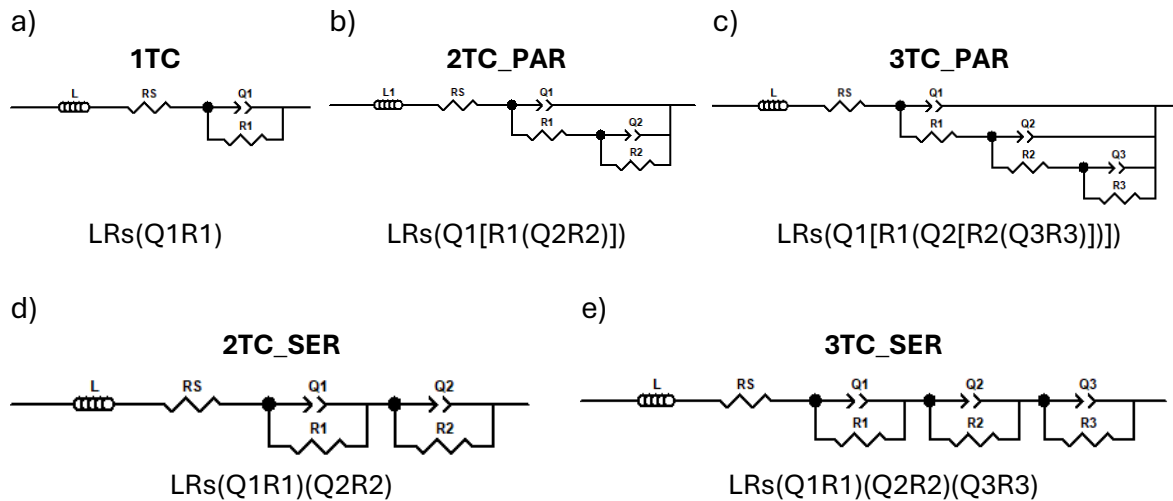


Figure 3. Equivalent-circuit topologies studied in this work. L: series inductance (high-frequency instrumentation/cable artefact); R_s : solution/electrolyte resistance; Q_n : constant-phase elements; R_n : charge-transfer/polarization resistances associated with each time constant (TC).

For the 1TC class, a single $R||Q$ element in series with the solution resistance (R_s) and an inductive element (L , accounting for high-frequency instrumentation/cable inductance), $LRs(Q1R1)$, was used (Figure 3a). L can be set to zero in the input spectrum if not needed, in which case the corresponding topology without an inductive element is used instead.

For the 2TC and 3TC classes, two topological arrangements of the additional time constants were tested, since the same number of time constants can, in principle, correspond to two physically distinct circuit configurations:

- Nested/parallel (PAR) topology in which each subsequent $R||Q$ element is nested within the resistive branch of the preceding one, $LRs(Q1[R1(Q2R2)])$ and $LRs(Q1[R1(Q2[R2(Q3R3)])])$, (Figure 3b,c). This configuration represents hierarchically coupled processes, where a given relaxation occurs "underneath" or in series with the resistive pathway of another.

- Series (SER) topology in which the R||Q elements are connected independently in series with one another, LRs(Q1R1)(Q2R2) and LRs(Q1R1)(Q2R2)(Q3R3), (Figure 3d,e). This configuration represents independent, non-interacting relaxation processes (classic Voigt-type circuit) [19], each associated with a physically separate interface or layer.

Both PAR and SER variants were independently CNLS-fitted to the same measured spectrum for the 2TC and 3TC cases, yielding five distinct reference parameter sets in total (1TC, 2TC_SER, 2TC_PAR, 3TC_SER, 3TC_PAR), given in Table S1 in the Supplement. Table S2 reports the corresponding Jacobian-based χ^2 reduced parameter uncertainty (% error). For each of these five fits it indicates how well each parameter is constrained by its measured spectrum.

For genuinely series (non-nested) topologies, in which the two RC-CPE branches are physically interchangeable, parameter sets were canonicalized by ordering the branches according to increasing relaxation time ($\tau = R \cdot Q^{(1/n)}$), following the same principle used elsewhere to resolve branch-swapping symmetry [20] (Leonardi et al., 2026), removing the resulting two-fold labelling ambiguity prior to spectrum simulation.

The synthetic reference database was constructed around five manual CNLS fits in Table S1, by perturbing reference parameter sets, one for each of the five topology variants. Around each reference parameter set, ten discrete perturbation levels were generated per parameter (five below and five above the reference value), using asymmetric, parameter-specific bounds. The complete set of parameter-specific perturbation bounds is given in Table 1.

Two complementary sampling strategies were applied: (i) single-parameter sweeps, in which each parameter was varied individually across its ten-point grid while all others were held at the reference value, generating local sensitivity trajectories, and (ii) randomized joint sampling ($n = 10000$ draws), in which each parameter was independently assigned one of its ten discrete perturbation levels, densely populating the local neighbourhood of the reference point in parameter space.

Table 1. Parameter-specific perturbation bounds used to construct the synthetic reference database. Bounds for R_s , L , and the CPE magnitudes/resistances (Q , R) are multiplicative, applied relative to the corresponding reference value in Table S1. Bounds for the CPE exponents n are additive and clipped to the physical range [0.5, 1.0].

Parameter	Perturbation range	Perturbation type
R_s	$\times 0.80 - \times 1.20$	multiplicative
L	$\times 0.70 - \times 1.30$	multiplicative

Parameter	Perturbation range	Perturbation type
Q1	$\times 0.30 - \times 3.00$	multiplicative
n1	± 0.10	additive
R1	$\times 0.30 - \times 3.00$	multiplicative
Q2	$\times 0.20 - \times 5.00$	multiplicative
n2	± 0.15	additive
R2	$\times 0.20 - \times 5.00$	multiplicative
Q3	$\times 0.20 - \times 5.00$	multiplicative
n3	± 0.15	additive
R3	$\times 0.20 - \times 5.00$	multiplicative

Impedance spectra for all synthetic parameter sets were then generated analytically via the corresponding equivalent-circuit transfer function, at the same frequency points as the original measured spectra, yielding $\log|Z|$ and phase angle for each entry in the database. Table 2 summarizes the number of synthetic spectra generated for each equivalent-circuit topology, broken down by generation method (systematic single-parameter variation and random joint-parameter combinations).

Table 2. Number of synthetic spectra generated per equivalent-circuit topology. P is the number of parameters perturbed for that topology (Rs, Q, n, R, and the series inductance L, varied consistently across all topologies). Each parameter contributes 10 systematically varied values (5 below and 5 above the reference value). Random variation denotes independently sampled joint parameter combinations.

Circuit	Number of parameters	Systematic variation (10×P)	Random variation	Total
1TC	5	50	10000	10051
2TC_SER	8	80	10000	10081
2TC_PAR	8	80	10000	10081
3TC_SER	11	110	10000	10111
3TC_PAR	11	110	10000	10111
Total				50435

Figure 4 shows the Bode representation of the systematic single-parameter variations comprising the synthetic reference database, for all five equivalent-circuit topologies (1TC, 2TC_SER, 2TC_PAR, 3TC_SER, 3TC_PAR), together with the corresponding PCA loading spectra. Panels (a) and (b) display the $\log|Z|$ and phase-angle spectra, respectively, generated by systematically perturbing each circuit parameter (R_s , Q_n , n , R_n , L) around its reference value. The resulting spectra span a broad but topology-consistent range of impedance magnitudes and phase-angle shapes across the studied frequency range.

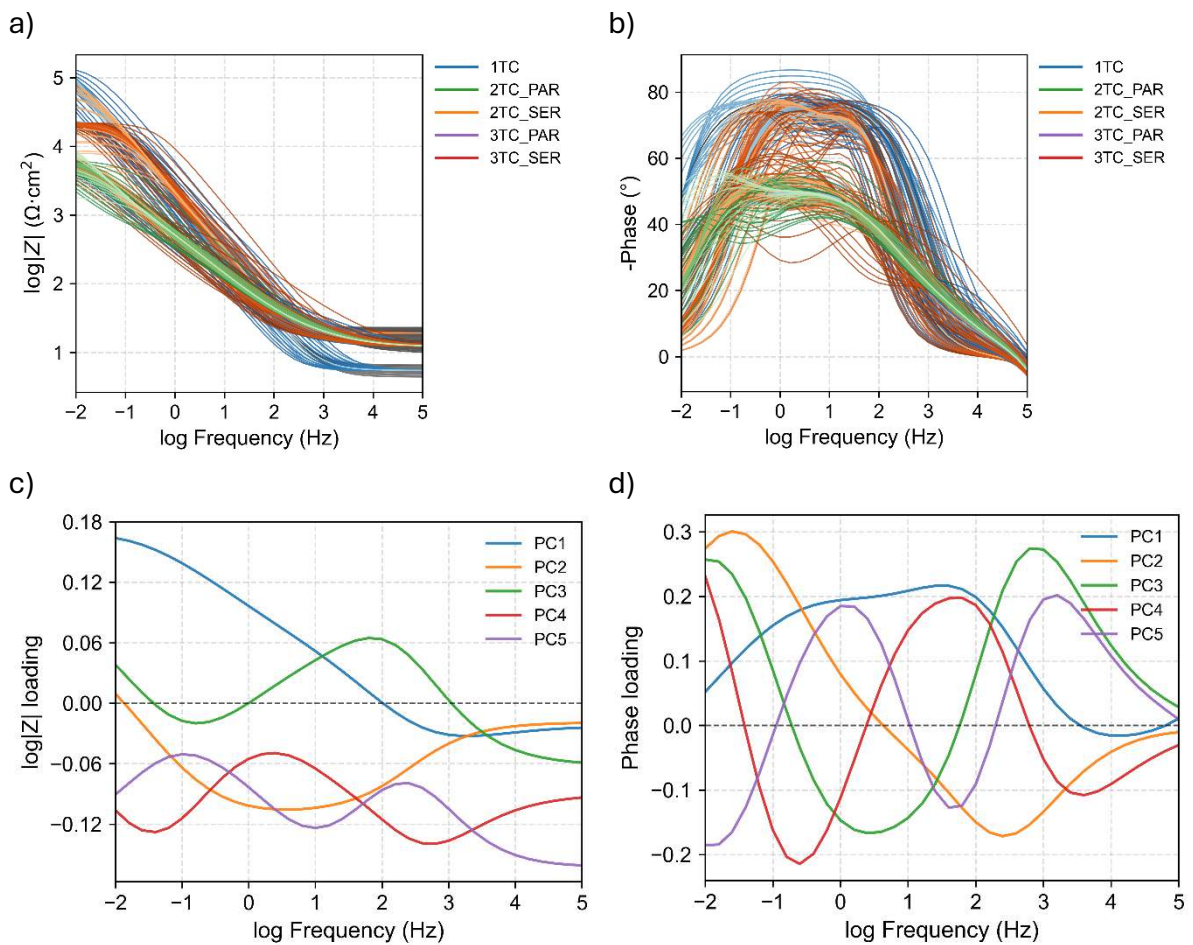


Figure 4. Bode plots of the systematic single-parameter variations in the synthetic reference database for the five studied equivalent-circuit topologies: $\log|Z|$ (a) and phase angle (b). PCA loading spectra for principal components PC1–PC5: $\log|Z|$ (c) and phase angle (d).

Panels (c) and (d) show the loading spectra of the first five principal components (PC1–PC5) obtained from PCA applied jointly to the $\log|Z|$ and phase-angle data. PC1–PC5

together account for 97.62% of the total variance in the dataset (individual contributions: PC1 = 43.69%, PC2 = 34.37%, PC3 = 13.60%, PC4 = 3.67%, PC5 = 2.29%) and were retained as the reduced representation used for subsequent classification and recovery.

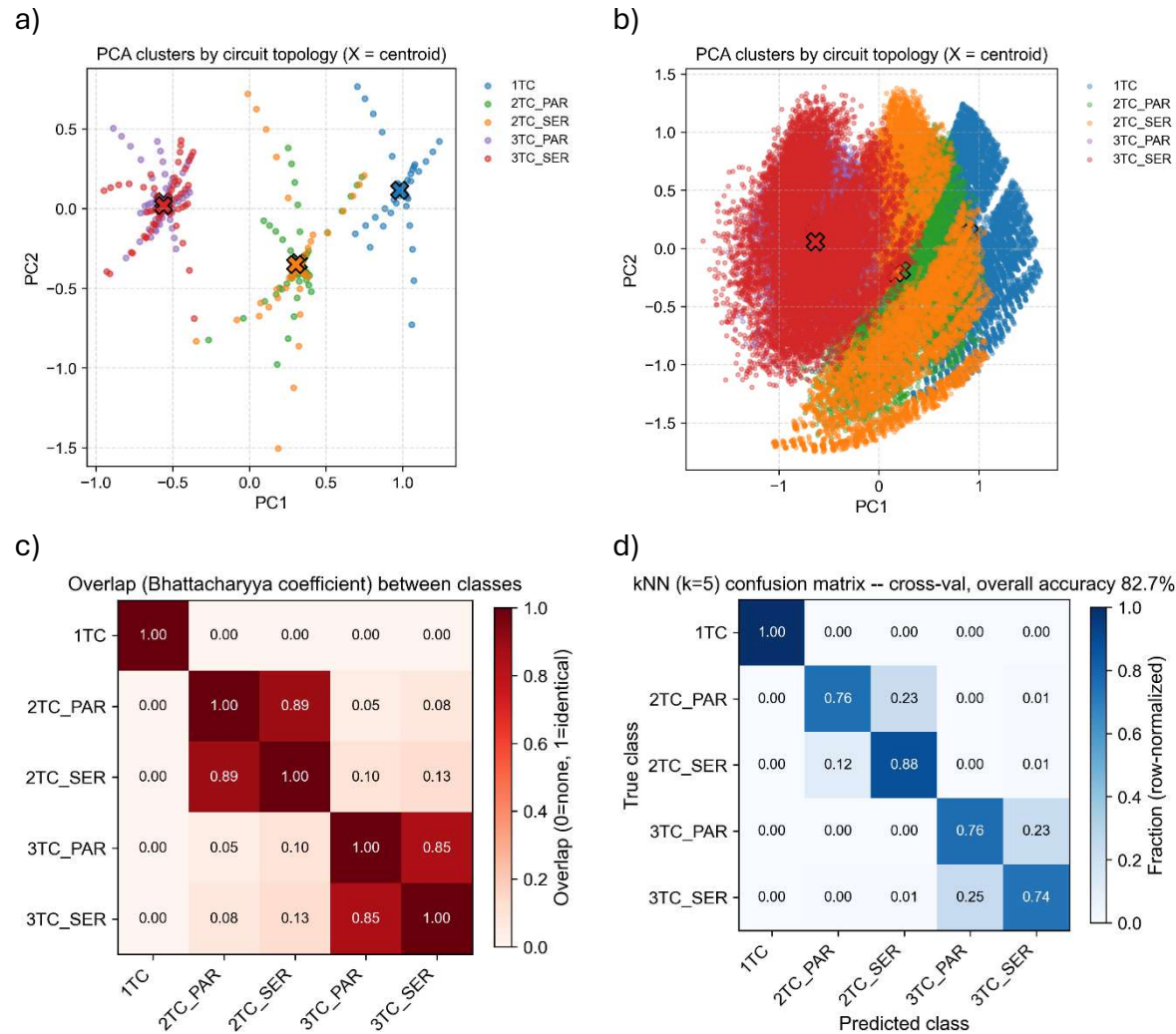


Figure 5. PCA clusters by circuit topology (X = centroid) for (a) the systematic single-parameter variations only and (b) the full synthetic database (systematic and random joint variations combined), together with (c) pairwise cluster overlap (Bhattacharyya coefficient) and (d) the empirical k-NN (k=5) cross-validated confusion matrix (overall accuracy 82.4%), both computed on the full database.

Figure 5 presents the class separability of the five equivalent-circuit topologies in PCA space and the corresponding classification performance. Panels (a) and (b) show the PCA score clusters (PC1 vs. PC2, with class centroids marked by X) for the systematic

single-parameter variations alone and for the full synthetic database (systematic and random joint variations combined), respectively. In both representations, the topologies separate primarily along the number of time constants: 1TC forms a distinct, well-isolated cluster, while the 2TC and 3TC classes occupy adjacent, partially overlapping regions of PCA space. Notably, the SER and PAR variants of the same time-constant count (2TC_SER/2TC_PAR and 3TC_SER/3TC_PAR) show substantial mutual overlap, reflecting their shared number of relaxation processes despite differing topological connectivity, consistent with prior reports that PCA struggles to separate structurally analogous circuit configurations, such as capacitor-to-CPE substitutions [21]. Panel (c) quantifies this overlap using the Bhattacharyya [22] coefficient computed on the full database, confirming near-total separation of 1TC from all other classes (coefficient = 0) and high overlap within each TC-count pair (0.90 for 2TC_par/2TC_ser and 0.81 for 3TC_par/3TC_ser), with comparatively minor cross-talk between the 2TC and 3TC groups. Panel (d) shows the corresponding empirical k-NN (k = 5) cross-validated confusion matrix [23], which achieved an overall accuracy of **82.7%**, with 1TC classified with 100% accuracy. The remaining confusion is confined almost entirely to distinguishing the SER and PAR variant of the same time-constant count (2TC_PAR/2TC_SER confused in 23%/12% of cases, and 3TC_PAR/3TC_SER in 23%/25% of cases), consistent with the high pairwise overlap reported in panel (c).

It should be noted here that this ambiguity is not detrimental to the fitting workflow: each spectrum still occupies a unique, well-defined point in PCA space regardless of which topology it represents, and the topology assignment itself is not fully fixed a priori but resolved post hoc during CNLS refinement. Mixing of right and wrong topologies at the PCA-NN stage therefore amounts, at worst, to a suboptimal initial guess, which is discarded post hoc.

In the limiting case of exact mathematical degeneracy between two topologies (i.e. where an algebraic parameter transformation exists such that both circuits produce identical impedance for every frequency), this degeneracy would enter the pipeline through the synthetic database itself: the two topologies' simulated spectra would coincide at every point spanned by this transformation, producing overlapping points in the PCA latent space, so that the nearest-neighbour search would identify both topologies as equally close candidates and pass both forward for CNLS refinement. Fitting these two distinct models to the same real measured spectrum can still yield a measurable χ^2 difference between them, but this difference is not inherently informative: it may simply reflect the asymmetry of numerical convergence (since neither optimizer path nor floating-point precision is perfectly symmetric between the two parametrizations), carrying no genuine physical significance, or it may instead reflect that one parametrization more naturally accommodates the true underlying physical process, with the other requiring a less physically plausible combination of parameter values to reach the same fit quality. Because χ^2 alone cannot distinguish a

numerical artefact from a genuine physical preference in this regime, the pipeline's post hoc χ^2 -based selection would, in such a case, report a "winner" without indicating which of these two explanations applies, demonstrating the point that resolving true model non-uniqueness requires independent physical evidence beyond the impedance spectrum itself [1].

Sanity check

A sanity check on the pipeline's internal consistency involved automated fitting of five manually fitted reference spectra, covering all tested circuit topologies (1TC, 2TC_SER, 2TC_PAR, 3TC_SER, 3TC_PAR). In every case, the pipeline identified an exact match to the corresponding reference circuit directly within the candidate database. These results confirm the internal consistency of the automated fitting pipeline across all tested circuit topologies.

Fitting of known measured spectra

The pipeline was then applied to the three actual measured reference spectra, the same measurements used to construct the synthetic reference database. Figure 6 visualizes the recovered NN+CNLS fit (red dashed) overlaid on the measured test spectrum (black circles) as Bode $|Z|$, Bode phase, and Nyquist plots for each of the three topologies, showing the quality of agreement across the full frequency range.

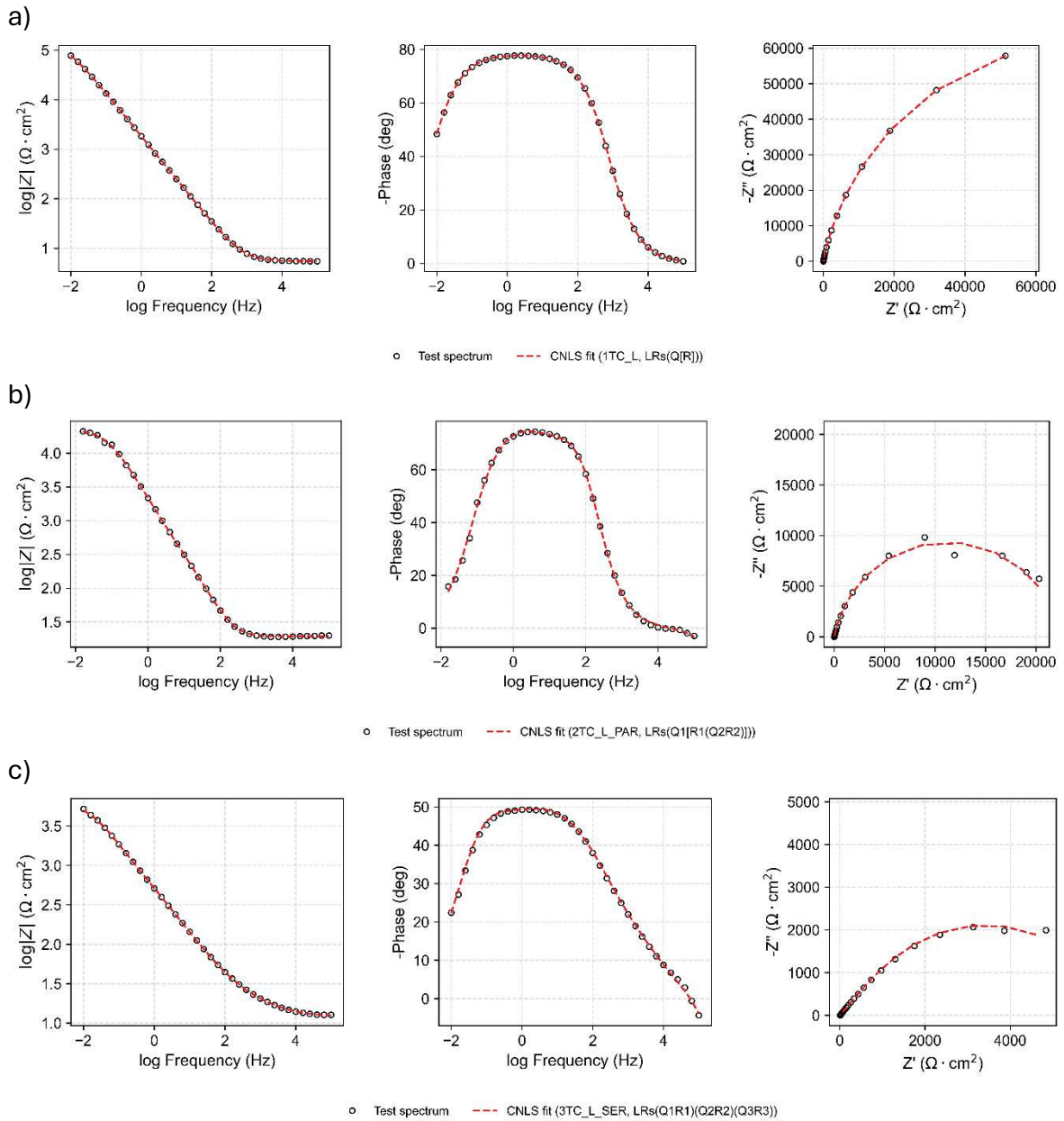


Figure 6. NN+CNLS fitted spectra (red dashed) against the measured test spectrum (black circles), with topology automatically determined by the pipeline, shown as Bode |Z|, Bode phase, and Nyquist plots, for (a) the 1TC reference circuit (LRs(Q[R])), (b) a 2TC parallel-with-inductor variant (LRs(Q1[R1(Q2R2)])), and (c) a 3TC series-with-inductor variant (LRs(Q1R1)(Q2R2)(Q3R3)).

The winning automated fits (Tables S3, S4 and S5) are compared against the original manual CNLS fits (Tables S1 and S2) to assess how well the fully automated workflow reproduces a result obtained by manual fitting. Tables S3, S4 and S5 also report the runner-up fits.

Firstly, across all measured spectra, the automated workflow identified a circuit topology matching the one selected by manual fitting (Table S1). For all the topologies the match was straightforward, as only a single topology was retained.

Secondly, the recovered parameter values are close to the corresponding manual fits across all three topologies, agreeing within 0.4% for 1TC, 0.7% for 2TC_PAR, and 0.9% for 3TC_SER (Tables S3–S5), and the reduced χ^2 is lower for the automated fit than for the manual fit in every case (1TC: 8×10^{-6} vs. 8×10^{-4} ; 2TC_PAR: 3.8×10^{-4} vs. 9×10^{-4} ; 3TC_SER: 8.8×10^{-5} vs. 2×10^{-4}). The Jacobian-based parameter uncertainties are likewise lower for the automated fit than for the corresponding manual fit. The sole exception is L in the 1TC case, where the automated fit converges to a value roughly two orders of magnitude smaller than the manual fit, together with a correspondingly larger relative uncertainty. This combination of a small point estimate paired with a large uncertainty indicates that L is not meaningfully resolved by the data in this frequency range, rather than indicating a genuinely small inductance. The objective function is essentially flat along the L-axis, so both the manual and automated fits converge to values that are statistically indistinguishable from one another despite their apparent 100-fold difference.

Fitting of unknown measured spectra

The synthetic reference database, by construction, is expected to provide reliable coverage only within the parameter envelope spanned by its seed spectra. This envelope-dependence is confirmed empirically. Two qualitatively distinct levels of distributional shift were examined for spectra not used to seed the reference database. The first, in-system generalization, concerns spectra termed “unknown”, measured under the same experimental conditions and instrumentation as the reference database, but originating from a different material than any of its seed spectra. These spectra represent a moderate shift, for which parameters are expected to fall close to, or within, the existing calibration envelope. The second, corresponding to the “genuinely unknown” spectra, concerns spectra originating from a substantially different measurement system altogether. These spectra represent a more severe shift, for which envelope coverage carries no such guarantee. For those spectra no manually fitted seed, and consequently no synthetic neighbourhood, existed in the database prior to this test. The manual CNLS fits of all spectra in this chapter were obtained independently, allowing the automated NN+CNLS fit to be evaluated as a genuinely blind prediction against an unrelated ground truth (manual fit).

Two representative examples of unknown spectra are shown in Figure 7. In case (a), the automated pipeline identified the spectrum as belonging to the 2TC_L_SER topology [LRs(Q1R1)(Q2R2)], while case (b) was identified as 3TC_L_PAR

[Rs(Q1[R1(Q2[R2(Q3R3)])))]). For unknown 2TC and 3TC spectra, where multiple topology candidates were viable, the matching topology was correctly identified as the one with the lowest reduced χ^2 . In both cases, the CNLS fit closely reproduces the measured $\log|Z|$, phase, and Nyquist representations across the full frequency range.

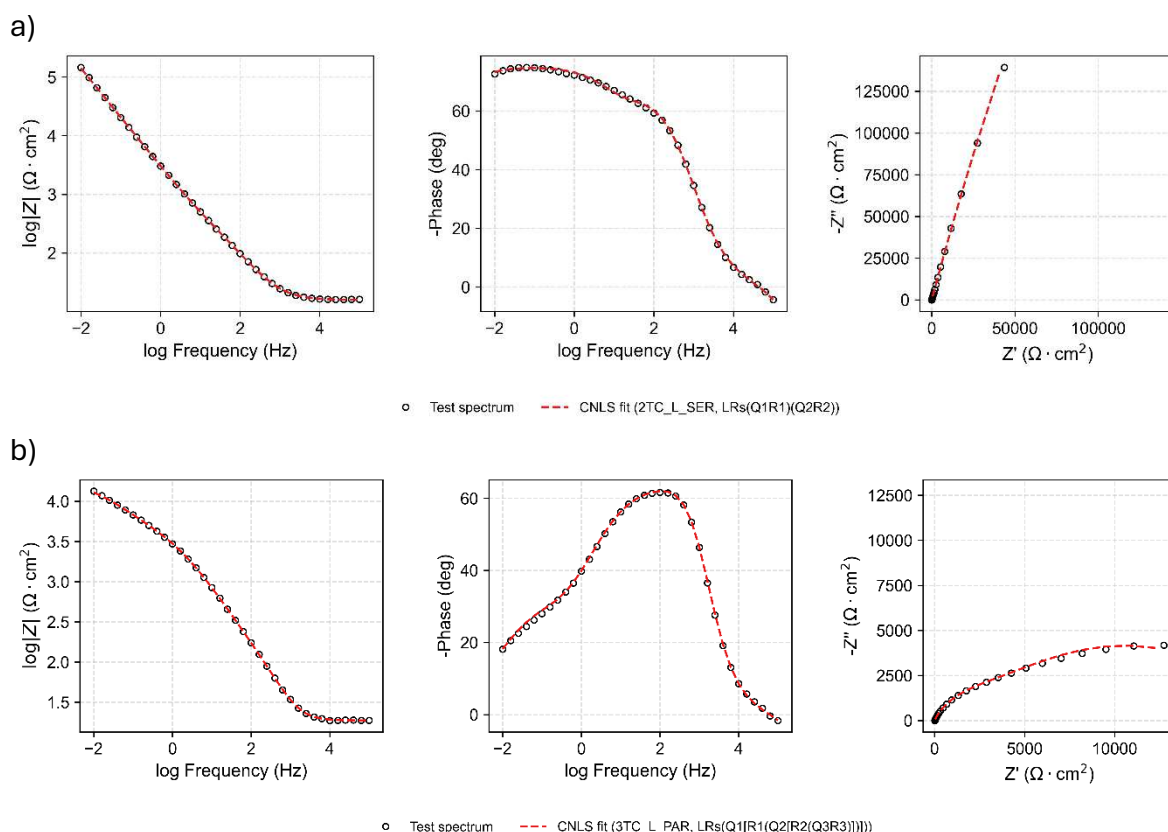


Figure 7. CNLS-refined fits (red dashed) against the measured test spectrum (black circles), shown as Bode $|Z|$, Bode phase, and Nyquist plots, for the best-performing candidate solutions: (a) 2TC, series with inductor (LRs(Q1R1)(Q2R2)) and (b) 3TC, parallel with inductor (LRs(Q1[R1(Q2[R2(Q3R3)])))).

To assess the reliability of these automated results beyond visual and χ^2 agreement, each spectrum was independently fitted by hand, and the two fits were compared parameter by parameter (Tables S6 and S7). For the 2TC_SER spectrum (Table S6), the manual and automated fits agree within 1% for every parameter, including R2, which, despite being poorly determined in both fits individually (manual error 48.81%, automated error 32.32%), converges to essentially the same value ($4.28 \times 10^6 \Omega$ vs. $4.237 \times 10^6 \Omega$) regardless of which fitting procedure is used. The automated pipeline also achieved a lower reduced χ^2 than the manual fit (0.000130 vs. 0.0003). For the 3TC_PAR spectrum (Table S7), the two fits agree within 0.74% for every parameter, including R1, which is poorly determined in both fits individually (manual error 34.3%, automated

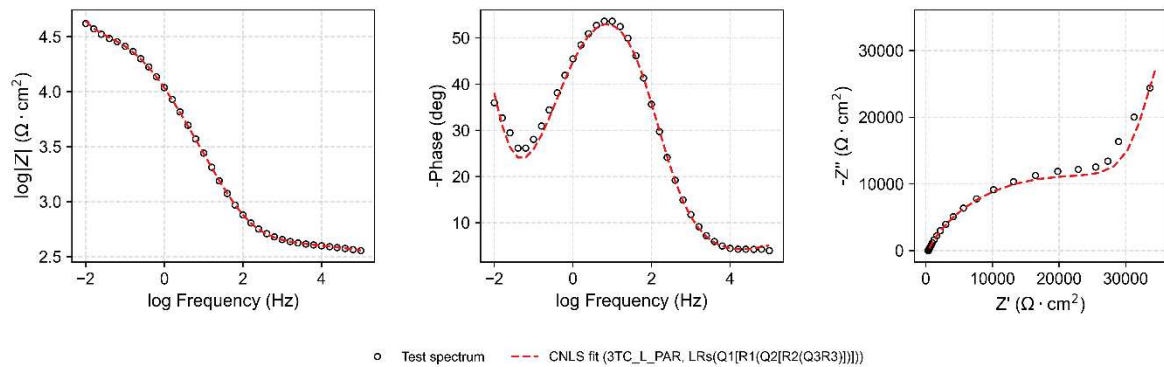
error 21.89%) yet converges to essentially the same value (192.5 Ω vs. 193.92 Ω). The automated fit again reached a lower reduced χ^2 (0.000093 vs. 0.0002).

Taken together, these two independent comparisons indicate that the automated NN+CNLS pipeline does not merely produce a numerically converged fit, but reproduces the same parameter set, including the same poorly determined, high-uncertainty parameters, that a careful manual analysis would yield.

Fitting of genuinely unknown measured spectra

The genuinely unknown spectra were deliberately chosen to probe the limits of the approach, surfacing three distinct challenges: an incompletely resolved low-frequency process, near-degenerate time constants, and parameter combinations falling outside the parameters and normalization envelope spanned by the reference database.

a)



b)

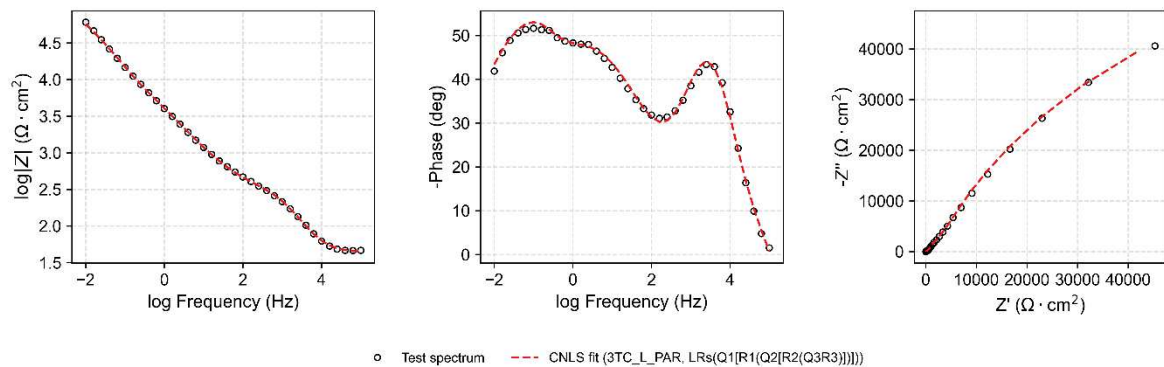


Figure 8. CNLS fits (red dashed) against two genuinely unknown measured test spectra (black circles), both within the [0,1] normalization envelope and both correctly identified as the 3TC_L_PAR topology $[LRs(Q1[R1(Q2[R2(Q3R3)]))]$, shown as Bode $|Z|$, Bode phase, and Nyquist plots. (a) Spectrum with a non-identifiable R3 (lower bound only, Table S6), χ^2 (reduced) = 0.000302, RMSPE = 1.80%. (b) Spectrum with near-

470 degenerate time constants (τ_2 and τ_3 closely spaced), χ^2 (reduced) = 0.000661,
471 RMSPE = 2.78%.

472 Fitting of genuinely unknown spectra originating from a substantially different
473 measurement systems, but still within the [0,1] normalization range, is shown in Figure
474 8. For the spectrum in Figure 8a, manual exploration established that the reduced χ^2
475 remains unchanged for any R_3 value at or above $8 \times 10^5 \Omega$. This value was therefore
476 adopted for the manual fit (Table S8). The automated fit, refining R_3 freely, converged to
477 a different, larger value ($2.88 \times 10^6 \Omega$) with a very large relative uncertainty (803.53%),
478 consistent with the same flat region of the objective function. The remaining
479 parameters (Q_1 , Q_2 , Q_3 , n_2 , n_3) agree within 10% between the two fits (Table S8),
480 confirming that only the non-identifiable R_3 , and the parameters correlated with it (R_s ,
481 n_1 , R_1), are affected. Similarly, as observed for L in Table S3, the data here establish a
482 lower bound on R_3 , and any value at or above roughly $8 \times 10^5 \Omega$ is equally consistent with
483 the measurement. This represents a limiting case of parameter non-identifiability. No
484 fitting procedure, manual or automated, can resolve a value the data themselves do not
485 constrain.

486 For the spectrum in Figure 8b, the manual and automated fits agree within 0.51% for
487 every parameter (Table S9), despite the spectrum containing two nearly overlapping
488 time constants ($\tau_2 \approx 3.05$ s and $\tau_3 \approx 3.88$ s, differing by only ~ 0.1 decade), which is
489 reflected in the comparatively large individual χ^2 -scaled uncertainties of the associated
490 parameters (R_2 , Q_3 , n_3) — even though both fitting procedures nonetheless converge
491 to essentially the same parameter values. This overlap is not merely a limitation of the
492 NN+CNLS approach, but near-degenerate time constants are a well-known source of
493 genuine non-uniqueness in CNLS fitting [2], where two or more R-CPE branches
494 become difficult or impossible to resolve individually regardless of the number of
495 starting points used.

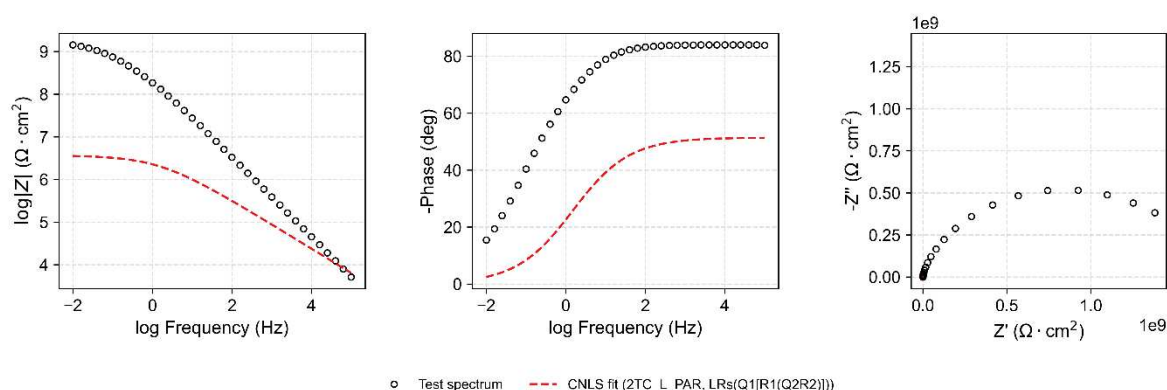
496 Establishing quantitative resolution limits for the two cases, i.e., the range of parameter
497 values below which a process remains unresolved by the measurement, and the
498 minimum time-constant separation, in decades, below which distinct topologies can
499 no longer be reliably distinguished by the present pipeline, would require careful and
500 systematic exploration, and is beyond the scope of this work.

501 Testing on a spectrum with markedly higher impedance (normalized $\log|Z| \approx 1.2$,
502 exceeding the [0,1] range established by the calibration set) confirmed this limitation
503 directly, leading to a completely wrong fit (Figure 9a). This confirms that reliable
504 automated fitting with this approach requires the reference database to be
505 representative of the system under study. The success of fit is a property of the
506 approach's coverage, not of the NN+CNLS mechanism itself: the database is
507 straightforwardly extensible, and adding a single manually fitted reference spectrum for
508 the new regime, together with its systematic and random synthetic neighbourhood,

would be expected to restore reliable performance for that class of spectra, without requiring any change to the existing database or pipeline.

To confirm this, the reference database was extended with two fitted spectra (2TC PAR and SER, without L parameter) obtained from manual fit of the spectrum in Figure 9a, together with its associated systematic (one-parameter) and random (combined) synthetic neighbourhood, following the same generation procedure used for the rest of the database. This increased the number of spectra in the database to 70,579. Re-running the pipeline on the same previously misfitted test spectrum, without any other change to the database or pipeline, yielded a fit correctly identified as the 2TC_PAR topology, matching the manually fitted reference, with a reduced χ^2 of 0.000265 and RMSPE of 1.84%. The two fits agree within 2.3% for every parameter except R_s , which is statistically undetermined in both fits, consistent with the negligible solution resistance expected for a coating-dominated system — even so, the two R_s point estimates differ by only 20% in absolute terms ($1.00 \times 10^{-6} \Omega$ vs. $8 \times 10^{-7} \Omega$) (Table S10, Figure 9b).

a)



b)

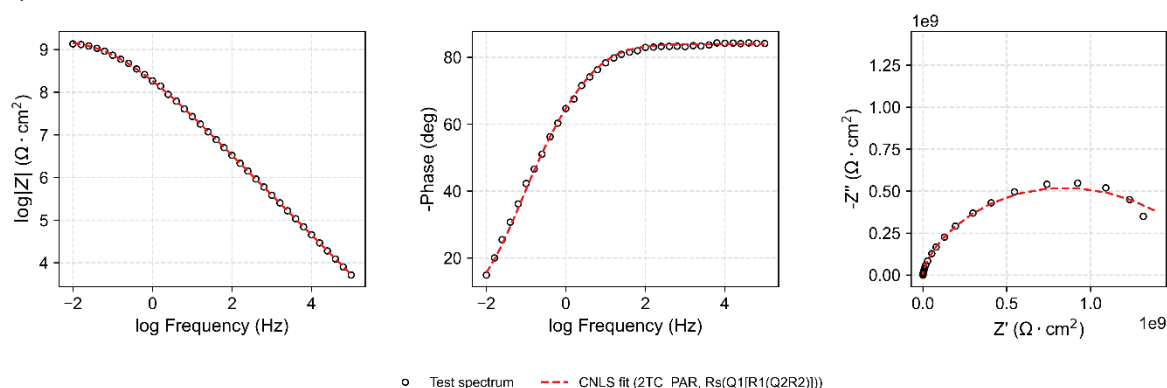


Figure 9 Improvement in CNLS fit stability and topology selection attained by extending the reference database with a manually fitted spectrum representative of a previously unstably-fitted measurement regime. (a) CNLS fit obtained prior to database extension, converging to the 2TC_L_ PAR topology [LRs(Q1[R1(Q2R2)])] and showing a pronounced

systematic deviation from the measured spectrum across $\log|Z|$, phase, and Nyquist representations. (b) CNLS fit obtained on the identical test spectrum after database extension, converging to the 2TC_PAR topology [Rs(Q1[R1(Q2R2)])] and closely reproducing the measured spectrum across all three representations, robustly across repeated database regenerations.

More importantly, the two additional reference spectra left every previously reported fit unchanged except that of the spectrum in Figure 7a where the correct 2TC_SER topology was entirely displaced from the candidate list at the default pool of 400 nearest neighbours, leaving 2TC_PAR as the only available candidate and yielding a fit with a still-low χ^2 but visibly different, less well-determined parameters. Only after increasing the neighbour pool from 400 to 2000 did 2TC_SER reappear among the retrieved candidates, restoring the results shown in Table S6.

This observation confirms that the imprecision was attributable to local database sparsity in that region of parameter space, rather than to inherent limitation of the workflow itself, and demonstrates that the pipeline can be incrementally extended to target a specific underrepresented topology. However, it also shows that it is favourable to build the reference database from spectra representative of the systems under study, since a distant, unrelated reference spectrum may destabilize fits of spectra with already poorly determined parameters and increase computational time.

Computational time

The pipeline involves three distinct computational stages, run independently of one another. Generating the full reference database for the seven seed circuits used in this work (five initial and two added during database extension) took 87.4 s in total. Fitting the PCA transformation on the resulting database is a separate, one-time step, repeated only when the reference database itself is extended, and took 3.9 s. Matching and CNLS-refining a given test spectrum against the already-transformed database is repeated for every spectrum fitted and typically completed in 5.6–6.6 s per spectrum at 400 nearest neighbours, with one exception: the spectrum in Figure 8a, whose R3 parameter was only weakly bounded from below (Table S8), took 11.1 s. Increasing the neighbour pool to 2000 raised the typical fitting time to 14 s, and to 28.6 s for the spectrum in Figure 8a.

Conclusions

This work demonstrates the operation of an automated equivalent-circuit fitting pipeline that treats topology selection and parameter initialization as a single problem

rather than two separate ones, and that avoids committing to a topology before evaluating how well it actually fits the measured spectrum. A synthetic reference database, built analytically from a handful of manually fitted seed spectra, was shown to jointly address both needs. PCA-space nearest-neighbour matching against this database simultaneously narrows the set of candidate topologies and supplies each with a physically informed starting point for CNLS refinement, with the final choice made only afterward, based on fit quality. Across three levels of validation, from internal consistency checks through known measurements to genuinely unseen spectra, the automated pipeline reproduced independent manual fits within a few percent for every parameter, including parameters that were poorly constrained in both the manual and automated fits individually, and in all cases achieved both a lower reduced chi-squared and lower parameter errors than the manual fit it was compared against.

The pipeline was shown to tolerate substantial overlap between structurally similar topologies in the reduced PCA representation, since a misleading nearest-neighbour match costs at most a discarded CNLS attempt, also shown in the pipeline output. This offers a practical way of living with the non-uniqueness inherent to equivalent-circuit fitting, in the spirit of recent calls for tools that surface competing candidates rather than force a single answer.

Three further points remain open for future work. First, near-degenerate time constants within a correctly identified topology, and low-frequency processes whose plateau falls outside the measured frequency window, both represent genuine physical limits on how well the associated individual parameters can be determined, ones that no amount of additional starting points or database coverage can overcome. Establishing quantitative resolution limits for this pipeline under each of these conditions would require a dedicated sensitivity analysis. Second, the size of the nearest-neighbour pool used for topology retrieval and the number of CNLS multistart points both proved to be practically important, tunable parameters rather than fixed settings: retrieval pools that were adequate for most spectra tested left the correct topology entirely absent from the candidate list for others, requiring an increase from 100 to 400, and in one case to 2000, neighbours to recover it. General guidance for choosing these values ahead of time, rather than by post hoc inspection as done here, would require further study. Third, the present reference database covers five equivalent circuit topologies, and extending it to a broader family of circuits would further test the generality of the approach demonstrated here. Ongoing work is extending this approach in three directions: replacing the manually fitted seed spectra with a fully synthetic reference database generated directly from theoretical circuit models, introducing a multistage refinement strategy in place of the single-stage CNLS step used here, and incorporating additional circuit topologies into the reference database to broaden the range of systems the pipeline can address.

605

606 **Methods**

607 Machine-learning workflows were implemented in Orange (version 3.32.24), an open-
608 source visual data mining platform. Built-in widgets were combined with custom
609 Python code through Python Script widgets to extend functionality. All workflow files
610 (.ows), documentation, and instructions for reproduction are provided in the data
611 repository (see Data Availability). The full Orange source code is openly accessible at
612 <https://github.com/biolab/orange3>. Manual fitting was done in PStace software by
613 PalmSens. Manual fit values and uncertainties are reported at the precision exported by
614 PStace; automated fit values are reported at full numerical precision from the fitting
615 routine.

616

617 **Code and data availability**

618 The datasets and code generated and analysed during the current study are available in
619 the Zenodo repository <https://zenodo.org/records/22922777> , comprising: the Orange
620 workflow used to generate the synthetic reference database from the manually fitted
621 seed spectra (Generate workflow.ows), the reference measured and manually fitted
622 spectra (Reference spectra) and the resulting generated database (pkl base), the
623 Orange workflow used for PCA-space nearest-neighbour matching and automated
624 CNLS fitting (Fitting workflow.ows), instructions for the use of both workflows, the
625 measured spectra used as fitting inputs (Input spectra), and, for every spectrum fitted
626 in this work, the complete fitting output showing all competing candidate topologies
627 side by side with their reduced χ^2 and parameter tables, not only the reported winner
628 (Fitting output).

629

630 **References**

- 631 1. Orazem, M. E. & Ulgut, B. Can interpretation of electrochemical impedance
632 spectroscopy data be automated? Where do artificial intelligence algorithms stand?
633 *Current Opinion in Electrochemistry* **55**, 101788 (2026).
- 634 2. Schaeffer, J. *et al.* Machine Learning Benchmarks for the Classification of Equivalent
635 Circuit Models from Electrochemical Impedance Spectra. *J. Electrochem. Soc.* **170**,
636 060512 (2023).
- 637 3. Nuñez Perez, F. A. Analytical–Computational Integration of Equivalent Circuit
638 Modeling, Hybrid Optimization, and Statistical Validation for Electrochemical
639 Impedance Spectroscopy. *Electrochem* **6**, 35 (2025).
- 640 4. Chen, W. *et al.* An intelligent matching method for the equivalent circuit of
641 electrochemical impedance spectroscopy based on Random Forest. *Journal of*
642 *Materials Science & Technology* **209**, 300–310 (2025).

5. Sáez-Pardo, F., Giner-Sanz, J. J. & Pérez-Herranz, V. Optimization of a convolutional neural network for system identification in electrochemical impedance spectroscopy. *Journal of Electroanalytical Chemistry* **1007**, 119893 (2026).
6. Zhu, S. et al. Equivalent circuit model recognition of electrochemical impedance spectroscopy via machine learning. *Journal of Electroanalytical Chemistry* **855**, 113627 (2019).
7. Sáez-Pardo, F., Giner-Sanz, J. J. & Pérez-Herranz, V. Equivalent Electrical Circuit recommendation for Electrochemical Impedance Spectroscopy: A benchmark of different Machine Learning algorithms. *Journal of Electroanalytical Chemistry* **976**, 118812 (2025).
8. Bongiorno, V., Gibbon, S., Michailidou, E. & Curioni, M. Exploring the use of machine learning for interpreting electrochemical impedance spectroscopy data: evaluation of the training dataset size. *Corrosion Science* **198**, 110119 (2022).
9. Bongiorno, V., Michailidou, E. & Curioni, M. Towards the automated interpretation of impedance spectra from organic coatings using neural networks. *Corrosion Science* **230**, 111922 (2024).
10. Chen, C., Jiang, X., Huang, K. & Xi, W. Auto-EIS: Automated Discovery of Equivalent Circuit Models from Impedance Spectroscopy via Deep Reinforcement Learning. in *2025 10th International Conference on Power and Renewable Energy (ICPRE)* 65–70 (IEEE, Hangzhou, China, 2025). doi:10.1109/ICPRE67300.2025.11274023.
11. Bongiorno, V., Hijnen, N. & Zhou, X. Exploring the use of large language models for EIS: A feasibility study on organic coatings for corrosion protection. *Corrosion Science* **261**, 113629 (2026).
12. Jaber, A. et al. AutoREC: A reinforcement learning platform for equivalent circuit model generation. Preprint at <https://doi.org/10.48550/arXiv.2604.27266> (2026).
13. Boukamp, B. A Nonlinear Least Squares Fit procedure for analysis of immittance data of electrochemical systems. *Solid State Ionics* **20**, 31–44 (1986).
14. Bandarenka, A. S. Development of hybrid algorithms for EIS data fitting. (2012).
15. Iki, R., Furugori, M. & Katayama, N. Automated Estimation of Equivalent Circuit Model from Impedances with Long Short-Term Memory.
16. Jolliffe, I. T. *Principal Component Analysis*. (Springer, 2002). doi:10.1007/b98835.
17. Demšar, J. et al. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* **14**, 2349–2353 (2013).
18. Martínez, S., Hudec, B., Martínez, I. & Šoić, I. Unsupervised clustering of EIS spectra for distinguishing degradation states of polymer coatings. *Progress in Organic Coatings* **215**, 110082 (2026).
19. Orazem, M. E. Measurement model for analysis of electrochemical impedance data. *J Solid State Electrochem* **28**, 1273–1289 (2024).
20. Leonardi, S. G. et al. Automated electrochemical impedance spectroscopy fitting for quantitative analysis of degradation modes in lithium-ion batteries. *Journal of Energy Storage* **154**, 121144 (2026).
21. Makogon, A., Kanoufi, F. & Shkirskiy, V. Is Unsupervised Dimensionality Reduction Sufficient to Decode the Complexities of Electrochemical Impedance Spectra? *ChemElectroChem* **11**, e202300738 (2024).

22. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. in (1943).
23. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**, 21–27 (1967).

Author contributions

Sanja Martinez: Writing – original draft, Methodology, Conceptualization, Supervision.
Lea Hrženjak: Investigation, Writing — review & editing.

Funding

This research was conducted within the project MXeneFun (112112) financed by the European Union's- NextGenerationEU fund from the source 581 – The recovery and resilience mechanism, in the frame of Programme financing of public higher education institutions and public scientific institutes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementASimpleModularPipelineforAutomatedEISEquivalentCircuitFittingviaPCASpaceNearestNeighbourMatchingValidatedAgainstManualFitting.do](#)