POLITECNICO DI TORINO

Master of Science in Georesources and Geoenergy Engineering
LM-35 (DM270)

Master's Degree Thesis

A Machine Learning Framework for Caprock Sealing Efficiency: Prediction of Breakthrough and Snap-off Pressure from Petrophysical Properties.

Supervisor

Prof. Vera Rocca

Candidate

Hichem Khalil Cherfouf

Academic Year 2025 / 2026

DEDICATION

First and foremost, all praise and gratitude belong to Almighty God for providing me with the strength, patience, and guidance to complete this journey.

To my wife, ***Ikram,*** *whose endless patience, strength, and devotion kept our family grounded across the distance while I pursued this master's degree. You carried the weight of our daily life alone so I could focus on building our future, and I am forever grateful for your love and sacrifice.*

To my darling daughter, ***Kounouz****, whom I had to leave when you were only five months old. Watching you grow through a screen was the hardest part of this journey, but every smile and laugh during our video calls was my greatest motivation. This thesis is as much yours as it is mine.*

I extend my deepest gratitude to my supervisor, ***Vera Rocca****, for her exceptional guidance, continuous support, and invaluable mentorship throughout my research.*

My sincere thanks go to all my professors and to all polito family for their dedication to my academic journey.

Finally, my heartfelt thanks go to my friend ***Karim*** *for his steadfast support, camaraderie, and encouragement during our time abroad together*

Abstract

The geological storage of carbon dioxide and hydrogen depends on a caprock's sealing efficiency, its capacity to retain a buoyant fluid over long timescales. Sealing efficiency is characterized jointly by two capillary pressures: the breakthrough pressure, at which the injected fluid first forces its way through the water-filled pore throats of the seal, and the snap-off pressure, the residual sealing capacity that remains once flow has occurred and the wetting phase re-imbibes. Measuring either takes weeks per plug, so the field relies on the Thomas (1968) correlation with permeability, and more recently on machine-learning models that report coefficients of determination approaching unity. Those models are validated by splitting a compiled dataset at random, which allows a model to be tested on rocks from studies it has already partly learned.

This thesis asks what can be known about caprock sealing pressure from routinely available petrophysical data, and tests prediction as it would actually be used: under source-aware validation, holding out whole studies rather than individual records.

A compilation of laboratory sealing-pressure measurements was traced record by record to its primary publication; nearly two thirds of it did not survive that check, the most consequential exclusion being a block of twenty-five records whose target values came from three different studies while their experimental conditions were placeholder values. The verified foundation — 146 records from 14 sources, every one carrying a DOI and a verification status — supports the two modelling populations analyzed in this thesis: permeability and porosity both correlate with breakthrough pressure (Spearman $\rho = -0.73$ and -0.76), porosity carries no comparable signal for snap-off pressure ($\rho = +0.08$, not significant), and a three-group data-driven seal-quality classification separates sealing behavior more sharply than lithology labels do.

Four models were compared under leave-one-source-out validation, in which each contributing study is withheld in turn, with three more flexible methods added as a check. Breakthrough pressure can be predicted with modest accuracy, and no model transfers better than a refitted permeability-only correlation, which reaches a coefficient of determination of 0.46 on studies it has never seen, within a median factor of 2.6 over the full range and of 1.9 on the rocks that would be assessed as caprocks. A random forest, a gradient-boosted ensemble, a support-vector regressor, an AdaBoost regressor and a neural network all fit their training data better and transfer no better; a two-feature linear regression that adds porosity matches the correlation's transferable coefficient without exceeding it, lowering the full-range median error from a factor of 2.6 to 2.0 but giving no reduction on the caprocks themselves. The refitted permeability-only law is therefore adopted as the estimator, with porosity retained only as a refinement in a chart. The record-level bootstrap intervals of both linear models exclude zero, while resampling whole studies leaves every interval spanning zero. That accuracy is well below the values the published literature reports, and the difference is traced to validation design and not to method. The learning curve was still rising when the data ran out, so additional independent sources and per-sample fluid–rock interaction measurements are the more promising next step, not more flexible models.

None of the methods tested predicts snap-off pressure from these inputs. Every leave-one-source-out coefficient is negative, and the likely reason is physical: snap-off is governed by the pore-body-to-throat aspect ratio, which bulk transport measurements do not encode. It should be bounded by the range measured within its seal-quality class and not predicted, until a pore-geometric measurement is available to enter as a feature.

The practical output is a tiered estimation framework that matches the route to the measurements available: a mercury-injection entry pressure converts to a breakthrough estimate through a fluid-conversion chart, to an order of magnitude and with a band of at least a factor of three; a permeability alone goes through the adopted correlation, whose transfer to an unseen study has been measured; and a permeability with a porosity is read from a chart carrying an in-sample band of about a factor of 2.4, which lowers the typical error to a factor of two. Each route states the accuracy it can support, and the verified, traceable dataset is offered as a benchmark for future work.

Keywords: caprock integrity, breakthrough pressure, snap-off pressure, CO_2 storage, machine learning, petrophysics.

Table of Contents

Abstract 3
Table of Contents 5
List of Figures 8
List of Tables 11
Nomenclature 13
Symbols 13
Statistics 13
Abbreviations 14
Populations used throughout 14
Chapter 1 — Introduction 15
1.1 Context and motivation 15
1.2 Problem statement 15
1.3 Aims and objectives 16
1.4 Contribution of this work 16
1.5 Scope 17
1.6 Thesis structure 18
Chapter 2 — Theoretical Background and Literature Review 19
Part I — The physics of caprock sealing 19
2.1 Caprocks in subsurface fluid storage 19
2.2 Caprock failure mechanisms 20
2.3 The capillary entry condition 21
2.4 Breakthrough pressure and snap-off pressure 22
2.5 Interfacial tension, contact angle, and fluid dependence 24
2.6 Petrophysical controls on sealing capacity 26
2.7 Pore-throat radius and rock-typing indices 28
2.8 Experimental measurement of sealing pressures 28
2.9 The Thomas (1968) correlation and its limitations 30
Part II — Methods for compiled data 31
2.10 Statistical methods for compiled datasets 31
2.11 Machine-learning concepts 32

2.12 Data-driven prediction of breakthrough pressure 33
2.13 Regression methods for small tabular data 34
2.14 Random forests and gradient boosting 35
2.15 k-means clustering and cluster validation 36
2.16 Source-aware validation 36
2.17 Research gap 37
Chapter 3 — Compilation, Verification and Analysis of the Dataset 39
Part I — Verification, enrichment and screening 39
3.1 The compilation as received 39
3.2 Verification against primary sources 40
3.3 Enrichment of the compilation 40
3.4 Screening criteria and their effect 42
3.5 The verified foundation 44
Part II — Analysis 48
3.6 Sealing pressures across fluid systems 48
3.7 Feature–target relationships 50
3.8 Outliers and their treatment 52
3.9 Within-study and between-study variance 55
3.10 Feature selection 56
3.11 Petrophysical grouping of the samples 57
3.12 Dataset limitations 60
Chapter 4 — Charts for Estimating Breakthrough Pressure 62
4.1 Chart BP-1: breakthrough pressure from permeability and porosity 62
4.2 Chart BP-2: fluid conversion and mercury injection 63
4.3 Validation of the mercury conversion 66
4.4 Application of the charts 67
Chapter 5 — Predictive Modelling under Source-Aware Validation 68
5.1 The prediction task 68
5.2 Models compared 69
5.3 The source-aware validation protocol 70
5.4 Breakthrough pressure results 72

5.5 Effect of model complexity 75
5.6 Feature contributions to transfer 76
5.7 Confining pressure as a feature 77
5.8 Effect of dataset size 79
5.9 Snap-off pressure results 80
5.10 Sensitivity to screening decisions 81
5.11 Summary of results 83
Chapter 6 — Discussion and Conclusions 85
6.1 Summary of the modelling results 85
6.2 Robustness to the removal of a source 86
6.3 Choice of accuracy measure 87
6.4 Accuracy by seal-quality class 88
6.5 Estimation from the charts alone 90
6.6 Chart and model compared 90
6.7 Limits of predictability 92
6.8 Comparison with published work 93
6.9 Gain over the Thomas correlation 95
6.10 Limitations 96
6.11 Recommendations for further work 97
6.12 Conclusions 98
References 100
Appendix A — Source publications and dataset provenance 104
A.1 Retained sources 104
A.2 Corrections applied during verification 105
A.3 Blocks removed and the grounds for removal 106
A.4 Sources identified but not retained 107
A.5 The delivered dataset 107
Appendix B — Supporting figures 109
B.1 The compilation and its completeness 109
B.2 The targets across fluids and rock types 111
B.3 The petrophysical grouping 114

List of Figures

Figure 2.1 — Conceptual model of caprock sealing: a buoyant plume accumulates beneath the seal and capillary menisci resist entry into the pore throats. Espinoza & Santamarina (2017).

Figure 2.2 — Capillary tube analogy. The non-wetting fluid enters a water-wet pore of radius r only once the pressure difference exceeds $2\sigma cos\theta / r$.

Figure 2.3 — Drainage and imbibition capillary-pressure curves for one sample. Breakthrough pressure is read on the drainage branch and snap-off pressure on the imbibition branch, the second lying below the first.

Figure 2.4 — Pore-scale controls on the two targets.

Figure 2.5 — Pressure signatures of the laboratory procedures.

Figure 2.6 — Under-fitting, appropriate fitting, and over-fitting of a model to noisy observations. The third panel passes through every point including its noise.

Figure 2.7 — Two tree-ensemble strategies. A random forest trains independent trees in parallel and averages them; gradient boosting adds trees in sequence, each fitted to the residual error of those before it.

Figure 2.8 — Random versus source-aware partitioning of a multi-source dataset.

Figure 3.1 — The merged compilation as received: by fluid and measurement method, and by lithology label and target.

Figure 3.2 — Records remaining after each screening step.

Figure 3.3 — Screening outcome and fluid composition of the verified foundation.

Figure 3.4 — Cumulative distributions of breakthrough and snap-off pressure.

Figure 3.5 — Records per source, with the cumulative share.

Figure 3.6 — Completeness of the verified foundation, by quantity.

Figure 3.7 — Mercury entry pressures as reported and after conversion to a common basis.

Figure 3.8 — Both pressures by fluid, with the ratio of medians.

Figure 3.9 — Permeability span within each lithology label.

Figure 3.10 — Feature–target correlations for the two pressures.

Figure 3.11 — Three outlier criteria applied to the same data.

Figure 3.12 — Breakthrough pressure against permeability and porosity.

Figure 3.13 — Snap-off pressure against the same two features. The porosity panel is flat.

Figure 3.14 — Snap-off relative to breakthrough pressure on the thirteen same-plug pairs.

Figure 3.15 — Silhouette score and smallest cluster size against the number of clusters.

Figure 3.16 — Chart SG-1: the three seal-quality groups in the porosity–permeability plane.

Figure 3.17 — Breakthrough and snap-off pressure by seal group.

Figure 4.1 — Chart BP-1: breakthrough pressure from permeability and porosity.

Figure 4.2 — Chart BP-2: breakthrough pressure for each storage fluid from a mercury-injection entry pressure. The mercury value is read on the bottom scale and the fluid line followed to the left scale; symbols are the same-plug checks.

Figure 5.1 — Random and source-aware partitioning, and the fold structure the design produces on the breakthrough population.

Figure 5.2 — The four model families. The Thomas baseline is one line in log-permeability, the multivariate model adds a porosity term, the forest averages independent trees and boosting corrects its own residual in sequence.

Figure 5.3 — Model comparison for breakthrough pressure: leave-one-source-out R^2 with 95 per cent record-level bootstrap intervals, and the same estimate against the random five-fold estimate.

Figure 5.4 — Predicted against measured breakthrough pressure, from leave-one-source-out predictions.

Figure 5.5 — The complexity stress-test. Each method is shown by its in-sample fit and its leave-one-source-out estimate, the span between them being the overfitting gap.

Figure 5.6 — Confining-pressure ablation, and how far the variable identifies its source study.

Figure 5.7 — Leave-one-source-out R^2 against the number of source studies used for training.

Figure 5.8 — Snap-off pressure: in-sample fit and leave-one-source-out estimate for each model.

Figure 5.9 — Transferable accuracy under alternative screening decisions.

Figure 6.1 — Leave-one-source-out accuracy with all eight sources and with each source withheld in turn (Table 6.1a).

Figure 6.2 — Restricting the population to better seals.

Figure 6.3 — Median prediction error by seal-quality class.

Figure 6.4 — The tiered estimation framework. Each route is matched to the measurements available and to the accuracy it can support; snap-off pressure is bounded by class and not predicted.

Figure 6.5 — Median error of both models across thresholds, and the residual bias by pressure range.

Figure B.1 — Contact angle and interfacial tension carried by each publication, by fluid.

Figure B.2 — Completeness by publication, as the percentage of each study's records reporting each quantity. The gaps form horizontal bands, one study at a time.

Figure B.3 — Confining pressure and test temperature against sample depth, with the in-situ gradients.

Figure B.4 — Both pressures by fluid, with the ratio of medians.

Figure B.5 — Sealing pressure by carbon-dioxide phase state, and the records in the pressure–temperature plane.

Figure B.6 — Breakthrough and snap-off pressure by lithology label, for labels with at least four records; counts in brackets.

Figure B.7 — Snap-off pressure by publication, showing the range of each study and its median.

Figure B.8 — Applied confining pressure by publication, with the 50 to 220 bar window shaded.

Figure B.9 — Silhouette of two partitions of the porosity–permeability plane.

Figure B.10 — Lithological composition of each seal-quality group.

List of Tables

Table 1.1 — Principal findings.

Table 2.1 — Laboratory procedures for characterizing caprock sealing pressure.

Table 2.2 — Data-driven studies of breakthrough pressure and related targets.

Table 2.3 — Property categories governing breakthrough pressure, and their use.

Table 3.1 — Studies added to the compilation.

Table 3.2 — Screening of the merged compilation.

Table 3.3 — Descriptive statistics of the two targets and the two features.

Table 3.4 — Kruskal–Wallis tests on the verified data.

Table 3.5 — Feature–target correlations.

Table 3.6 — Records flagged by three outlier criteria.

Table 3.7 — Between-study share of variance.

Table 3.8 — Cluster diagnostics against the number of groups.

Table 3.9 — The three seal-quality groups.

Table 4.1 — Breakthrough pressure read from Chart BP-1, in bar, at four porosities.

Table 4.2 — Mercury-to-gas conversion factors (Equation 4.2, $\vartheta_{gas} = 0°$).

Table 4.3 — Same-plug checks of the mercury conversion in the compilation (Equation 4.2, $\vartheta_{gas} = 0°$).

Table 5.1 — Model configurations.

Table 5.2 — Model comparison for breakthrough pressure (n = 38, eight sources, 29 plugs). R^2 computed on $\log_{10}P$; the three intervals are 95 % bootstrap percentile intervals on the leave-one-source-out estimate, resampling records, plugs and whole studies respectively.

Table 5.3 — Complexity stress-test, all seven methods under leave-one-source-out validation.

Table 5.4 — Confining-pressure ablation, separating the feature from the population.

Table 5.5 — Model comparison for snap-off pressure (n = 48, nine sources, 42 plugs). The interval is the 95 % record-level bootstrap interval.

Table 5.6 — Transferable accuracy under alternative screening decisions (multivariate linear, leave-one-source-out).

Table 6.1a — Leave-one-source-out accuracy with each source withheld in turn.

Table 6.1 — Model performance by breakthrough-pressure threshold (multivariate linear, leave-one-source-out).

Table 6.2 — Prediction accuracy by seal-quality class, from leave-one-source-out folds.

Table 6.3 — The tiered estimation framework: route, evidence and stated accuracy.

Table 6.4 — Class-specific snap-off bound, and the spread of measured snap-off about it.

Table 6.5 — Median prediction error, Thomas baseline against the multivariate model, leave-one-source-out.

Table A.1 — Source publications of the verified dataset.

Table A.2 — Value corrections, each made against the published table.

Table A.3 — Blocks excluded from the verified foundation.

Nomenclature

Symbols

Symbol	*Quantity*	*Unit*
P_b	*breakthrough pressure*	*bar*
P_{snap}	*snap-off pressure*	*bar*
P_{Hg}	*mercury-injection entry pressure*	*bar*
P_c	*capillary pressure*	*bar*
P_{conf}	*confining pressure*	*bar*
$\hat{P}$	*predicted pressure*	*bar*
k	*permeability*	*mD*
φ	*porosity*	*–*
r	*pore-throat radius*	*m*
d^*	*smallest pore diameter on a percolating path*	*m*
σ	*interfacial tension*	*mN/m*
θ	*contact angle*	*°*
CF	*mercury-to-gas conversion factor*	*–*
ψ	*pore shape factor*	*–*
T	*temperature*	*°C*
ε	*log-ratio prediction error*	*–*
a, b, c, d	*fitted regression coefficients*	*–*

Statistics

Symbol	*Quantity*
R^2	*coefficient of determination, computed on $\log_{10}P$*
RMSE	*root-mean-square error, in $\log_{10}$ units*
r	*Pearson correlation coefficient*
ρ	*Spearman rank correlation coefficient*
H	*Kruskal–Wallis test statistic*
p	*probability value*
SD	*standard deviation*
CV	*coefficient of variation*
IQR	*interquartile range*
n	*number of records*
k *(clustering)*	*number of clusters*

Abbreviations

Term	*Meaning*
CCS	*carbon capture and storage*
UHS	*underground hydrogen storage*
MICP	*mercury injection capillary pressure*
MIP	*mercury injection porosimetry*
SbyS	*step-by-step method*
LOSO	*leave-one-source-out cross-validation*
S1	*the breakthrough-pressure modelling population*
S2	*the snap-off-pressure modelling population*
ARI	*Adjusted Rand Index*
DOI	*digital object identifier*

Populations used throughout

Name	*n*	*Sources*	*Plugs*	*Definition*
merged compilation	*413*	—	—	*as received plus enrichment*
non-duplicate	*348*	—	—	*after duplicate removal*
verified foundation	*146*	*14*	*98*	*after screening (Table 3.2)*
S1 breakthrough	*38*	*8*	*29*	*flow-through records with both features*
S2 snap-off	*48*	*9*	*42*	*residual records with both features*
mercury set	*40*	*3*	—	*routed to Chapter 4*

Chapter 1 — Introduction

1.1 Context and motivation

The geological underground fluid storage is central to climate change mitigation. Carbon dioxide captured from industry must be placed in deep formations and kept there for centuries. Hydrogen produced when renewable electricity is abundant must be stored at scale and recovered when it is scarce, and natural gas has long been stored seasonally by the same means. In every case the fluid is injected into a porous reservoir and must not escape upward. In every case the barrier that prevents escape is the same: a caprock, a layer of fine-grained rock whose pores are too narrow for the buoyant fluid to enter.

The property that decides whether a caprock holds is its breakthrough pressure, the pressure at which the injected fluid first forces its way through the water-filled pore throats of the seal. Below that pressure the seal is intact. Above it, the fluid migrates and containment is lost. A related quantity, the snap-off pressure, describes the residual sealing capacity once flow has occurred and the wetting water re-imbibes. Together these two capillary pressures characterize how well a formation will contain a stored fluid, and estimating them reliably is a prerequisite for any storage project.

1.2 Problem statement

Breakthrough pressure is measured in the laboratory on core samples, but the measurement is slow and demanding. A step-by-step test raises the pressure on a brine-saturated plug in increments until flow is detected, and a single measurement can take weeks. Because such data are scarce, the field has long relied on a shortcut: the Thomas correlation (Thomas et al., 1968; a three-author paper, cited here in the customary short form), which estimates breakthrough pressure from permeability alone through a simple power law. It is fast and it captures a first-order trend. But it uses only one of the several properties the capillary physics identifies as controlling, and its accuracy on rocks outside its original calibration has not been established.

More recently, machine-learning studies have predicted breakthrough pressure from several petrophysical properties and reported high accuracies, with coefficients of determination approaching unity. These results are promising, but they share a feature that this thesis will show to be consequential. They are validated by splitting a compiled dataset at random, which allows a model to be tested on rocks from studies it has already partly learned. Whether such a model would transfer to a genuinely new site is a different and harder question that the reported figures do not answer.

A second problem sits beneath the first, and this thesis found it to be larger than expected. A compilation assembled from the literature carries the errors of its assembly: values entered in the wrong column, permeabilities derived and not measured, records whose cited source contains no such measurement, and blocks whose experimental conditions are placeholders. None of this is visible in a summary statistic. All of it changes what a model learns.

1.3 Aims and objectives

The aim of this thesis is to establish what can be known about caprock sealing pressure from routinely available petrophysical data, and to measure how far that knowledge transfers. Prediction is therefore tested under source-aware validation, in which whole contributing studies are withheld in turn, rather than by the random split of pooled records on which the published figures rest.

The objectives are:

- to assemble a compilation of laboratory sealing-pressure measurements, to verify every record against its primary publication, and to record that verification in the dataset itself so that any reader can repeat it;
- to characterize the two sealing pressures and their relationships with the petrophysical properties; through the correlation and outlier analyses of Chapter 3, a decomposition of each quantity into its within-study and between-study parts, and a data-driven grouping of the samples into seal-quality classes, and so to determine which properties the data can support as predictors;
- to test whether breakthrough and snap-off pressure can be predicted from those properties, using source-aware validation that measures transfer to an unseen study, and to compare a range of models from the Thomas correlation to machine-learning ensembles;
- to provide a practical means of estimating sealing pressure that states the accuracy it can support and the conditions under which it applies.

The thesis treats these as questions to be answered by the data, and reports the answers whether or not they favor the more elaborate method.

1.4 Contribution of this work

This work delivers three primary contributions: a verified and traceable benchmark dataset; a quantitative measurement of the cross-study transferability of data-driven sealing pressure estimates and a practical framework for operational estimation.

The first is the dataset. The compilation on which the thesis rests was traced record by record to its primary publication and checked value by value. What that check found is reported in Chapter 3 as a result in its own right rather than as a preliminary to the modelling, because any statement about how far a model transfers is conditional on the records being what they claim to be. Every retained record carries a DOI and a verification status, and every specimen a plug identifier, so that a reader can repeat the check. The dataset is delivered with the thesis and offered as a benchmark on which future methods can be compared on equal terms.

The second is the measurement of transfer. Four models, from the classical permeability correlation to two machine-learning ensembles, are compared under leave-one-source-out validation, with three more flexible methods added as a check, so that every prediction is made on rocks from a study the model has never encountered. Chapter 5 reports what that validation design costs relative

to the random-split figures the literature reports, and Chapter 6 draws the boundary of predictability that follows from it; which sealing quantity can be estimated, from which measurements, and to what accuracy.

The third is the framework itself: a tiered set of estimation routes matched to the measurements a field assessment is likely to have, each carrying an explicit statement of the accuracy it can support, together with two charts that give a first estimate by hand.

Table 1.1 — The questions this thesis asks, and where each is answered.

Question	***Answer***
Does the compilation support modelling as received?	*Sections 3.2 to 3.4*
Is breakthrough pressure predictable on a study the model has not seen?	*Sections 5.4, 6.1*
Does a more flexible learner do better than the classical correlation?	*Sections 5.5, 6.9*
Which petrophysical property carries the transferable signal?	*Sections 5.6, 6.8*
Can confining pressure be used as a feature?	*Sections 3.10, 5.7*
Is snap-off pressure predictable from the same inputs?	*Sections 5.9, 6.7*
What accuracy can a practitioner claim, and on what basis?	*Sections 6.3, 6.4, 6.6*

1.5 Scope

The work is bounded in four ways, and each bound rests on evidence the later chapters set out rather than on convenience.

Scale. The thesis concerns plug-scale laboratory measurements and not the field-scale behavior of a whole seal. A caprock at field scale is governed additionally by faults, fractures and lateral continuity, which core data do not capture, and the leak-rate literature reviewed in Section 2.1 finds leakage at most sites dominated by those pathways rather than by transport through the intact matrix. A plug measurement therefore characterizes a crack-free upper bound on sealing capacity, and that is the quantity modelled here.

Fluids. The compilation covers carbon dioxide, nitrogen, methane and hydrogen. Hydrogen coverage is thin: three records from one study enter the modelling population, and a second hydrogen study is present in the dataset but excluded because it does not report the porosity and permeability a model would need. When the compilation was first assembled, no core-scale hydrogen breakthrough measurement on brine-saturated plugs with accompanying petrophysics had been published. Two such studies appeared during this work; both are in the dataset, and only one supplies the petrophysics a model needs. Any hydrogen result here is therefore indicative and not established, and on the one study that tested hydrogen and nitrogen on the same plugs, hydrogen broke through at the lower pressure (Section 3.3).

Properties, and what is left out. The models use the two properties the compilation records at the sample level, permeability and porosity. The other properties the capillary physics identifies as

controlling are excluded on evidence, not by oversight. The fluid–rock interaction enters the Young–Laplace relation through the interfacial tension and the contact angle, but this compilation carries both only as study-level literature constants, a single value applied to every sample of a study, so neither can inform a model at the sample level (Sections 3.1 and 5.1); the fluid effect is handled after the rock term, as the conversion of Chapter 4, and Section 3.6 shows why it cannot be recovered as a feature, the threefold difference between fluids being lost in a rock-to-rock variation spanning seven orders of magnitude in permeability. In-situ conditions are set aside on the same kind of evidence. Confining pressure is recorded but behaves in this compilation almost as a source label, five of the six studies reporting an independent value applying one value to all their samples, and Section 5.7 shows that the accuracy it appears to add cannot be separated from study identity. Temperature acts on sealing through the interfacial tension, which the dataset carries only as a per-study constant, and the literature does not agree on the sign of its effect (Section 3.10). The carbon-dioxide phase state, liquid against supercritical, is a real physical dimension of the data but not a dominant control across these samples (Section 3.6). Depth carries no usable relationship to the laboratory conditions the tests were actually run at (Section 3.10). Water saturation, which the literature shows raises breakthrough pressure exponentially, is recorded by no source in the compilation. The absence of pore-geometric and fluid–rock interaction properties is not only a limitation but one of the findings, and the most likely reason snap-off pressure resists prediction at all (Section 6.7).

Screening conventions. Throughout, a formation with a breakthrough pressure below about 10 bar is treated as not a caprock, and one below roughly 20 to 30 bar as a poor seal. These are working conventions adopted in this study and not universal cut-offs. The 10-bar figure corresponds, at a representative brine–carbon dioxide density contrast, to only a short retainable column, so a rock below it cannot hold a column of practical height. The bands are used to interpret results and not to select data: low-value records are retained as training data so that the models span the full range of sealing behavior, and they are not presented as caprocks. Whether such a rock seals in the field is a column-height and scale-up question rather than a plug-scale one.

1.6 Thesis structure

Chapter 2 sets out the physics of caprock sealing: the capillary control on the two pressures, the properties that govern them, the laboratory methods that measure them, and the Thomas correlation, together with the statistical and machine-learning methods used later. Chapter 3 assembles, verifies and screens the dataset, characterizes the two targets, and derives the seal-quality classes used throughout. Chapter 4 develops two charts that estimate breakthrough pressure without a model. Chapter 5 fits and compares the predictive models under source-aware validation. Chapter 6 draws the results together into the boundary of predictability and the tiered framework, compares them with the literature, and states the limitations and recommendations.

Chapter 2 — Theoretical Background and Literature Review

This chapter sets out the physical and methodological ground on which the thesis rests. Sections 2.1 to 2.9 cover the physics of caprock sealing; the capillary control on breakthrough and snap-off pressure, the fluid and rock properties that govern it, the laboratory procedures that measure it, and the Thomas (1968) correlation that serves as the baseline. Sections 2.10 to 2.16 cover the statistical and machine-learning methods used in Chapters 3 and 5. Section 2.17 states the gap the thesis addresses.

Part I — The physics of caprock sealing

2.1 Caprocks in subsurface fluid storage

The geological storage of fluids depends on a low-permeability formation, the caprock, that overlies the reservoir and prevents the buoyant injected fluid from escaping upward. The injected phase is lighter than the resident brine, so it accumulates beneath the seal. It develops a buoyancy pressure proportional to the height of the column and the density contrast between brine and stored fluid (Espinoza & Santamarina, 2017). Whether that fluid remains trapped over the timescales required for storage is governed, in the absence of faults and fractures, by the capillary resistance of the intact caprock.

Most seals considered for storage are shales, mudstones, or evaporites. Their sealing behavior comes from the small size of their pore throats. Mechanical compaction, chemical diagenesis, and cementation during burial reduce both porosity and pore size. Shales typically show porosities of about one to twelve per cent, mean pore sizes between roughly five and one hundred nanometers, and permeabilities on the order of 10^{-21} to 10^{-19} m² (Espinoza & Santamarina, 2017). Pores of this size sustain capillary pressures of several megapascals, which is what allows a meter-scale seal to hold a substantial fluid column.

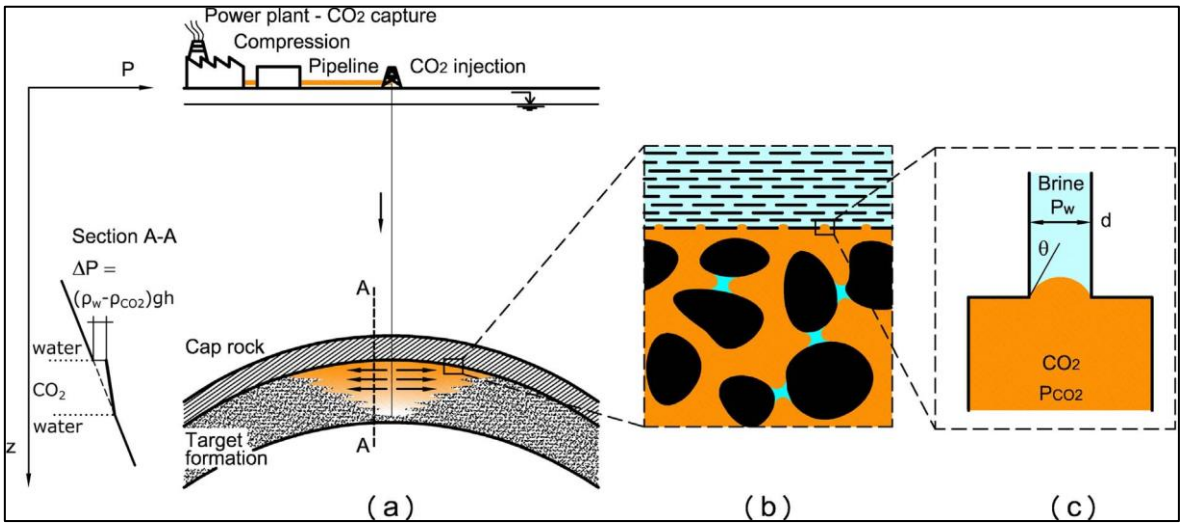


Figure 2.1 — Conceptual model of caprock sealing: a buoyant plume accumulates beneath the seal and capillary menisci resist entry into the pore throats. Espinoza & Santamarina (2017).

A striking consequence of this pore structure is how little of it carries the flow. Hildenbrand et al. (2002) determined the transport porosity of the conducting pore system in a suite of mudstones and clays to be between 10^{-5} and 10^{-2} per cent, against total porosities of 23 to 30 per cent. Gas moves through a vanishingly small fraction of the pore space, along paths whose equivalent radii they estimated at 8 to 60 nanometers. This is why bulk porosity and bulk permeability are indirect descriptors of sealing behavior, a point that recurs throughout this chapter.

Leak-rate analyses show that transport through a continuous, unfractured caprock is minor at most sites, leakage being dominated instead by faults, fractures and abandoned wellbores (Espinoza & Santamarina, 2017). Laboratory measurements on intact plugs therefore characterize a crack-free upper bound on sealing capacity, and that sample-scale property is what the models in later chapters predict.

2.2 Caprock failure mechanisms

A caprock can lose its sealing function through several distinct mechanisms. Espinoza & Santamarina (2017) group the degradation pathways into capillary invasion of the intact matrix, hydraulic fracture and fault reactivation under reservoir overpressure, and geochemical alteration as acidified brine dissolves and re-precipitates minerals.

Fractures have the largest reported effect. Kim et al. (2025), working on Opalinus Clay, found that fractures raised permeability by two orders of magnitude and reduced breakthrough pressure by a factor of five. Kim & Santamarina (2013) report a larger effect on shale plugs from a carbon-storage pilot site. Their three specimens gave breakthrough pressures of 10 to 30 kPa, which correspond to pore radii of several micrometers. That is far larger than the nanometer radii expected for those shales, and microscopy revealed hairline cracks parallel to bedding in every plug. After sealing treatment with bentonite and kaolin slurries the same plugs held shut-off pressures of 0.55 to 1.25 MPa, a factor of roughly twenty to sixty higher (Kim & Santamarina, 2013, Table 1). This result is the basis on which such measurements are excluded from the dataset of Chapter 3.

Repeated exposure can degrade the seal, as three studies report, while a fourth, with nitrogen, does not. Wollenweber et al. (2010) found breakthrough pressure falling by a third to a half over successive carbon-dioxide cycles on the same plugs, with permeability rising, and attribute it to carbonate dissolution and re-precipitation. Tonnet et al. (2011) observed effective permeability rising from 0.3 to 1.5 nanodarcy over successive resaturation cycles. Jia et al. (2026) quantified it most closely: breakthrough pressure fell from 4.83 to 2.17 MPa over 168 hours of exposure, a reduction of 55 per cent, with the mechanism confirmed as pore enlargement followed by plugging. Hildenbrand et al. (2002), by contrast, found repeat experiments reproducible with nitrogen, so the distinction appears to be chemical reactivity and not the repetition itself.

Wettability alteration is a further mechanism specific to carbon dioxide. Chiquet et al. (2007) found the cosine of the drainage contact angle on mica falling from about 0.9 at low pressure to about 0.5 at 11 MPa, attributing the shift to the pH drop as carbon dioxide dissolves, which cancels the surface

charge stabilizing the water film. Because the effect is strongest on mica and organic-rich surfaces, it separates the sealing of carbon dioxide from that of less reactive gases.

Of these mechanisms, capillary breakthrough of the intact matrix is the one measured by the procedures reviewed here and modeled in the chapters that follow.

2.3 The capillary entry condition

The pressure a non-wetting fluid must overcome to enter a water-saturated pore follows from the Young–Laplace equation. For a cylindrical pore of radius r, the capillary pressure is

$$P_c = \frac{2\sigma\cos\theta}{r} \qquad (2.1)$$

where σ is the interfacial tension between the two fluids and θ is the contact angle measured through the wetting phase (Rocca, 2021). Espinoza & Santamarina (2017) write the same relation more generally as

$$P_c = \frac{\psi\,\sigma\cos\theta}{d^*} \qquad (2.2)$$

where d^* is the smallest pore diameter along a percolating path and ψ is a shape factor. Both forms express the same physical content. Entry pressure rises as interfacial tension and the cosine of the contact angle increase, and as the controlling pore size falls.

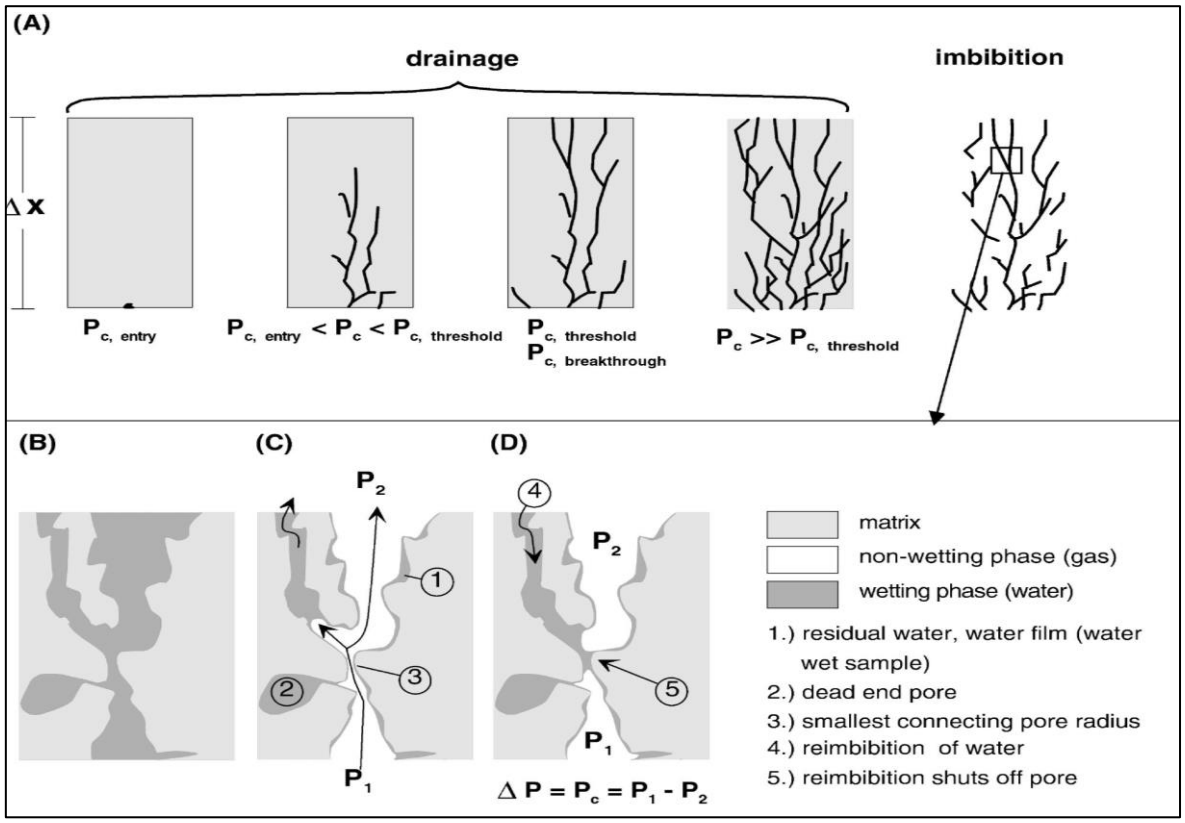


Figure 2.2 — Capillary tube analogy. The non-wetting fluid enters a water-wet pore of radius r only once the pressure difference exceeds $2\sigma\cos\theta/r$.

Mean pore size systematically over-predicts breakthrough pressure, because percolation occurs along paths made up of the larger pores and not the average pore. Measured values are therefore lower than those estimated from average pore size, and the gap widens for broad or multi-modal distributions. This is one reason a single geometric parameter rarely predicts breakthrough pressure across different rocks.

2.4 Breakthrough pressure and snap-off pressure

Breakthrough pressure is the pressure at which the non-wetting fluid first forms a connected path across a water-saturated sample during drainage. It is the point on the drainage capillary-pressure curve at which the invading phase percolates through, and it is obtained experimentally by the step-by-step procedure.

Snap-off pressure is the residual capillary pressure that remains once the wetting phase re-imbibes and disconnects the non-wetting phase in the largest pore throats. It is a point read from the imbibition branch of the capillary-pressure curve, obtained by the residual-pressure procedure, in which the decay of an initially applied pressure difference is tracked until it stabilizes.

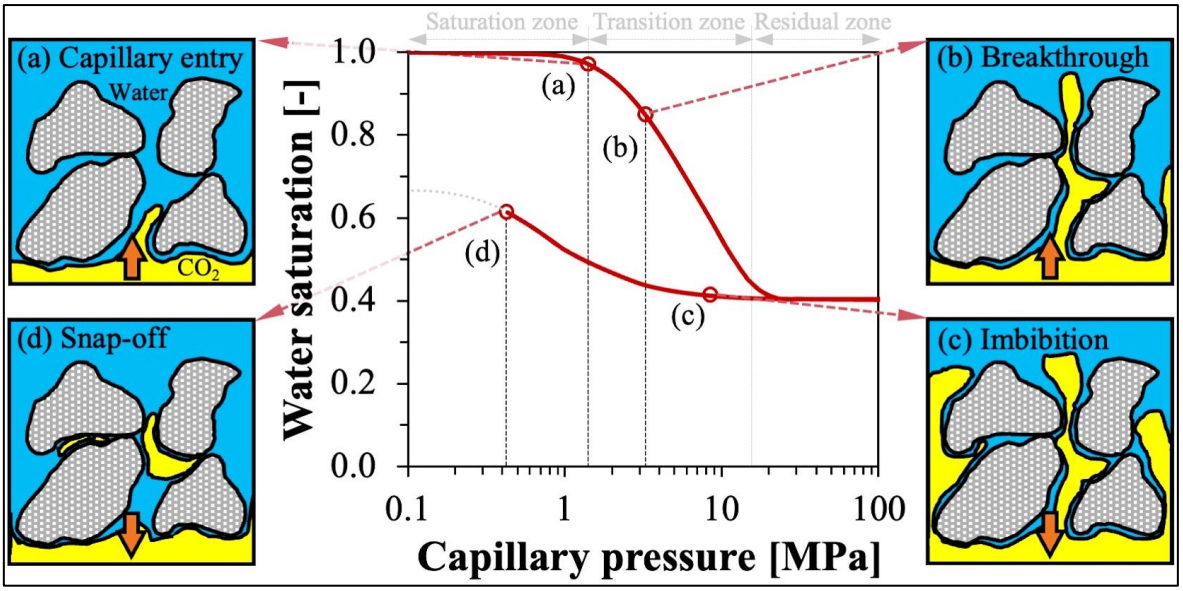


Figure 2.3 — Drainage and imbibition capillary-pressure curves for one sample. Breakthrough pressure is read on the drainage branch and snap-off pressure on the imbibition branch, the second lying below the first.

The consistent empirical finding is that snap-off pressure is lower than breakthrough pressure for the same sample. Kim & Santamarina (2013) describe the mechanism: after breakthrough the wetting phase re-imbibes, closing the smallest pores first, and the residual pressure at which the non-wetting phase stops flowing is set by the last connected pore-throat to shut off along the percolating path. That gives a lower value than the drainage breakthrough pressure and therefore a conservative, lower-bound estimate of sealing capacity.

Reported ratios of snap-off to breakthrough pressure vary more widely than a single range suggests. Ito et al. (2011) found the residual method giving 50 to 90 per cent of the step-by-step value.

Egermann et al. (2006) found a factor of 1.3 to 4.5, that is 22 to 77 per cent, depending on lithotype (computed from their Table 2, ambient air–brine tests). Boulin et al. (2013) and Amann et al. (2011) each found about half. A review by Chen et al. (2024) cites a 20 to 50 per cent range, at the lower end of what these primary measurements span, which is roughly 20 to 90 per cent. Kim et al. (2025) describe the imbibition method as conservative because of hysteretic capillary mechanisms.

The ratio is better treated as a benchmark than as a conversion factor, and Chapter 3 shows why. On the thirteen plugs in the present dataset carrying both measurements, the ratio ranges from 0.22 to 0.92 with a median of 0.48, inside the range the literature reports, and individual plugs differ by a factor of about four. A pooled ratio of medians compares different rocks measured in different studies. A same-plug ratio holds the rock fixed and changes only the procedure.

The terminology is not settled. The residual-method quantity is called snap-off pressure by several authors on the grounds that it reflects the pore-scale snap-off of the non-wetting phase during imbibition. Vespo et al. (2024), working numerically, argue that the residual capillary pressure measured in these tests should not be equated with a true pore-scale snap-off pressure, and state that it is not a property of the material. Boulin et al. (2013) likewise note that the snap-off process is acknowledged to differ from the entry-pressure process. The label “snap-off pressure” is used throughout this thesis for the residual-method quantity, consistent with the terminology of the source studies from which the data are drawn, while acknowledging that the residual measurement is an operational quantity whose micro-scale interpretation is still debated.

A practical consequence of the unsettled terminology appears in the data. Gao et al. (2014) describe their measured quantity as a carbon-dioxide breakthrough pressure while using the pulse-decay residual approach. Chapter 3 reclassifies those records to the snap-off target on the basis of the method used and not the name given.

The two pressures are set by different features of the same pore network. Breakthrough is a drainage-percolation threshold, fixed by the radius of the narrowest constriction along the widest connected path. That is a property of the pore-throat size distribution, which porosity and permeability constrain. Snap-off is an imbibition instability, and it depends not on throat size alone but on the ratio of pore-body radius to pore-throat radius. Roof (1970) showed that a wetting collar occupying a pore throat becomes unstable and disconnects the non-wetting phase only once that aspect ratio exceeds a critical value, of order 1.5 to 2 for idealized pore geometries. Lenormand et al. (1983) confirmed the mechanism experimentally in micromodels.

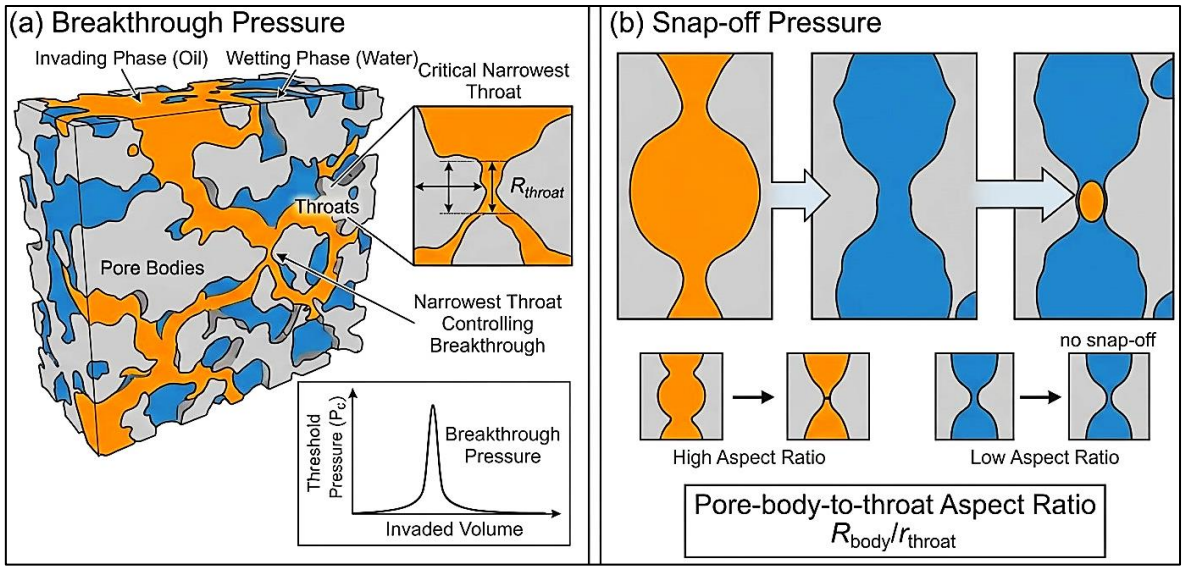


Figure 2.4 — Pore-scale controls on the two targets.

The pore-body-to-throat aspect ratio is a geometric attribute distinct from porosity and permeability. Two rocks with identical bulk properties can differ substantially in it, and it is not recoverable from a measured breakthrough pressure, which fixes the throat radius but carries no information about the pore-to-throat ratio above it. This decoupling is why breakthrough and snap-off are treated as separate quantities that are never inferred from one another.

2.5 Interfacial tension, contact angle, and fluid dependence

Because entry pressure scales with $\sigma\cos\theta$, the interfacial tension and wettability of the specific fluid pair control how sealing a given rock is to a given fluid.

For the carbon dioxide and brine system Bachu & Bennion (2009), from 378 pendant-drop measurements, express interfacial tension as

$$\mathrm{IFT} = A_{(T,S)}\, P^{-B_{(T,S)}} \qquad (2.3)$$

a power-law decline in which both the coefficient and the exponent depend on temperature and salinity. It falls steeply below the critical point, stabilizes near 20 to 25 mN/m once carbon dioxide is supercritical, rises with salinity, and dips sharply near the critical temperature.

The salinity dependence has a methodological consequence. Applying the correction requires a salinity value, which the compiled dataset lacks, so the correction cannot be applied rigorously. A constant reference interfacial tension is used only where a conversion between fluid systems is unavoidable, in placing mercury entry pressures on a gas–brine basis.

Contact angle varies with mineralogy and with fluid pressure, and the literature does not agree on how much. Chiquet et al. (2007) find the drainage contact angle on mica shifting substantially with carbon-dioxide pressure, reaching about 60° at 11 MPa. Fleury et al. (2011) report the opposite, that contact angles for the drainage process were barely affected by an increase in pressure to 14 MPa, and conclude that dense carbon dioxide does not significantly alter water-wettability in drainage.

The two papers share an author. The disagreement is unresolved, and it bears on the analysis here: Section 2.9 and Chapter 4 both depend on which value is adopted.

There is some independent support for the higher figure. Chiquet et al. (2007) back-calculate contact angles from the published breakthrough measurements of Hildenbrand et al. (2004) and Li et al. (2005), two of the sources in the present dataset, and recover cosines of about 0.4 to 0.6, which they describe as upper bounds. Those correspond to angles near 60°, which is the value the compilation carries for those records; the source studies themselves assumed 0° in their pore-size calculations.

The choice has a large effect on the result. Skerbisch et al. (2025), working in the Vienna Basin, computed the maximum retainable carbon-dioxide column height from one set of seal data under two wettability assumptions. A fixed water-wet angle gave more than 1000 metres at depths beyond 3000 metres. Allowing the angle to shift with depth gave a column peaking near 170 metres and falling to zero below about 2200 metres, which they describe as the absence of a static capillary seal beyond that depth. The same seal is either excellent or absent depending on the assumption.

Fluid ordering. The frequently reported ordering of breakthrough pressure on a given rock is $N_2 > CH_4 > CO_2$, which follows from the interfacial-tension ordering. Li et al. (2005) demonstrated the proportionality directly on Weyburn seal rocks, where the ratio of nitrogen to carbon-dioxide breakthrough pressure, 2.6 to 3.0, tracked the ratio of their interfacial tensions, 2.7 to 2.8. Their conclusion is that breakthrough pressure for a seal rock is nearly proportional to the interfacial tension of the gas and brine system.

The ordering is not universal: two studies report the relation failing on single plugs. Kawaura et al. (2013) measured supercritical carbon dioxide at roughly twice the value recalculated from their own nitrogen measurement through the Laplace relation. They suggest either that published carbon-dioxide interfacial tension is underestimated, or that interfacial tension and contact angle have a smaller influence on threshold pressure than the equation implies. Tonnet et al. (2011) found the opposite failure on a Rousse caprock plug. Nitrogen required 4.5 to 6 MPa while carbon dioxide broke through the same plug below 1.5 MPa. That is a factor of more than three where the interfacial tensions differ by less than two, with no wettability alteration detected. They describe the result as contrary to expectation. Taken together, these two studies show the conversion failing in both directions, and Chapter 4 uses them to justify the width of its uncertainty band.

Hydrogen. The compilation as originally assembled contained no hydrogen measurement, and three recent studies change that picture.

Benlalam et al. (2026) report nine step-by-step threshold-pressure experiments with hydrogen on a clay-rich caprock, with one methane test, and compare them against nitrogen data previously acquired on the same caprock and made available as a reference by Snam-Stogit. Their finding is that the three gases give comparable breakthrough pressures, with no evidence that sealing capacity falls when hydrogen is the displacing gas. Li, Gates & Hejazi (2025) measured hydrogen, methane and nitrogen breakthrough on the same outcrop shales and measured the interfacial tensions

directly: 71.57 mN/m for hydrogen, 72.11 for nitrogen and 72.03 for methane, within a per cent of one another. Their breakthrough pressures do not sit quite as close: on each of their three plugs hydrogen broke through 14 to 19 per cent below nitrogen, while methane was within 5 per cent of it (Section 3.3).

The wettability evidence points the same way. Ali et al. (2021) measured hydrogen contact angles on mica consistently lower than carbon dioxide, 49.7° against 83.5° at 323 K and 20 MPa (their §3.1 reports 42.9° for the same condition, an unresolved internal discrepancy). By Equation 2.1 a lower contact angle implies greater capillary resistance, so a caprock should seal hydrogen better than carbon dioxide on the wettability term. They note a caveat: at high organic-acid concentration and 25 MPa mica became hydrogen-wet at 98.8°, above the threshold at which capillary leakage becomes possible. Al-Mukainah et al. (2022) found the same organic control, a high-organic shale becoming fully hydrogen-wet while a low-organic shale stayed water-wet, and computed maximum trapped hydrogen columns of about 325 and 100 meters respectively. Their hydrogen–brine interfacial tension, 51 to 64 mN/m, sits below the other two studies, so a real spread remains in the literature.

2.6 Petrophysical controls on sealing capacity

The capillary relation ties entry pressure to pore size, interfacial tension, and contact angle. Pore size is not measured routinely, whereas porosity and permeability are. Much of the experimental literature therefore examines how breakthrough pressure relates to the measurable petrophysical properties that stand in for pore geometry.

Permeability is the property most consistently correlated with breakthrough pressure, because it integrates the size and connectivity of pore throats. Jia et al. (2026) found it the dominant control in Ordos caprocks. Hao et al. (2021), compiling 172 data sets (158 mudstone and shale, 14 sandstone) from sixteen publications, reported

$$\log P = -0.279 \log k - 5.097 \qquad (2.4)$$

with an overall coefficient of determination of 0.49 across 139 records. That is a real but scattered relationship, and its scatter is discussed below.

The correlation can also fail. Makhnenko et al. (2017) state that breakthrough pressure cannot be correlated with the intrinsic permeability of the caprock, and their evidence is a reversal. Intact Opalinus Clay is less permeable than remolded shale, 5×10^{-21} against 9×10^{-21} m², yet has a breakthrough pressure three times lower, 1.6 against 4.6 MPa. They attribute the inversion to microstructure, the intact shale containing channels with larger-than-dominant pore diameter. Benlalam et al. (2026) report the same on their samples: permeability was measurable on only three of their plugs, all below 1.6 nanodarcy, with no correlation to the measured breakthrough pressures possible. Chapter 6 returns to it.

Porosity relates to breakthrough pressure more ambiguously, and the ambiguity is instructive. Espinoza & Santamarina (2017) report breakthrough pressure increasing as porosity decreases and

specific surface increases, consistent with tighter, finer rocks sealing better. Hao et al. (2021) found that porosity's correlation was strong within a single data source but nearly absent, below 0.1, once data from many sources were pooled.

Two studies reviewed by Chen et al. (2024) report the opposite sign, breakthrough pressure increasing with porosity. One explanation, not tested here, is a sampling effect. Where a sample suite spans a compaction and diagenesis trend, porosity and throat size fall together and the correlation is negative. Where it spans a lithofacies contrast, clay-rich rocks carry their porosity in micropores with very small throats, so a rock can be more porous yet seal better. Kim, Vilarrasa & Makhnenko (2025) document the throat control directly: in their formations the clayey facies seal better than the sandy facies at comparable porosity (effective porosity is little changed between the two, while the sandy pore throats are up to thirty times larger), and they report breakthrough pressure to be correlated with clay content and pore-throat size but essentially independent of porosity. Whether porosity retains a signal in the present compilation is tested directly in Chapter3.

Pore-throat size and specific surface area are the properties most directly tied to the capillary relation. Where they are measured, they correlate strongly, and in the compilation of Hao et al. (2021) the maximum throat radius and permeability were the two properties most closely correlated with breakthrough pressure, at 0.57 and 0.49 respectively. They are rarely available across large multi-source datasets, which limits their use as model inputs.

Water saturation raises breakthrough pressure sharply, and no source in the present compilation records it. Zhang & Yu (2016) found breakthrough pressure rising exponentially with water saturation on Carboniferous shales, with the rise steepening above about 60 per cent. Zhao & Yu (2018) fitted the same exponential form on Ordos sandstone. Jia et al. (2026) fitted a power law. All three agree on the direction and on its size, which is larger than the fluid effect.

Confining pressure raises breakthrough pressure, because higher effective stress narrows the conducting throats. Kim et al. (2025) report a positive correlation between effective mean stress and breakthrough pressure. Tonnet et al. (2011) show how strong the stress dependence can be, permeability falling from 20 to 25 nanodarcy at 3.5 MPa effective stress to below 1 nanodarcy at 13.5 MPa on the same sample. The effect is real. Whether a compilation recording one stress value per study can resolve it is a separate question, taken up in Chapters 3 and 5.

Temperature has no agreed sign. Liu et al. (2025) measured hydrogen, methane and carbon dioxide on tight caprock at three temperatures and found breakthrough pressure falling linearly with temperature for all three gases, with coefficients of determination between 0.96 and 0.99. Jia et al. (2026), working on comparable rocks, found it rising, attributing the increase to partial evaporation of formation water generating additional pore pressure. Two careful studies in the same basin disagree on the direction of the effect.

Sample length is an uncontrolled variable. Jia et al. (2026) found breakthrough pressure scaling linearly with core length, with a coefficient of determination of 0.99. No source in the present compilation reports it.

2.7 Pore-throat radius and rock-typing indices

Where pore-throat radius cannot be measured directly, petroleum engineering has developed indirect estimates from the two properties that are routinely available. The Winland R35 relation, popularized by Kolodzie (1980), estimates the pore-throat radius at 35 per cent mercury saturation from a power-law regression on porosity and permeability, calibrated on reservoir sandstones. Pittman (1992) extended the approach with saturation-specific variants. Amaefule et al. (1993) took a related approach, combining porosity and permeability into a single flow-oriented index built on the Kozeny–Carman formulation.

The logic behind both schemes is the same. Pore-throat radius, and by extension flow and capillary behavior, can be inferred from the porosity and permeability pair without a direct pore-scale measurement. The limitation is also the same for a sealing problem. Both were calibrated on reservoir-quality sandstones for the purpose of predicting flow capacity, and both collapse porosity and permeability into a single index. A single flow-oriented quantity cannot distinguish a porous, low-permeability clay from a tight, low-porosity evaporite. The two rock types occupy comparable positions on a flow-based index but seal very differently, because flow capacity and capillary sealing capacity are not the same property. This is the specific reason Chapter 3 keeps porosity and permeability as separate inputs, and Chapter 5 puts the point to the test by carrying a Winland-index estimate as a baseline against the two features kept apart (Section 5.5).

2.8 Experimental measurement of sealing pressures

Several distinct laboratory procedures are used to characterize caprock sealing. Because they interrogate different points on the capillary-pressure curve, they return different numbers on the same sample. Mixing values from different methods into a single target would blur two physically distinct quantities.

The step-by-step method raises the applied gas pressure in discrete increments across a water-saturated sample until a continuous non-wetting path forms and flow is detected downstream. It measures breakthrough pressure on the drainage branch.

The residual-pressure method applies an initial pressure difference and tracks its decay as the wetting phase re-imbibes, until the residual capillary pressure stabilizes. This residual value is the snap-off pressure.

The dynamic method infers an entry pressure from flow measurements at several imposed rates, extrapolated to zero flow, and is faster than the step-by-step approach.

Mercury injection measures an entry pressure indirectly by injecting mercury into a dry sample and converting the result to the fluid pair of interest through the ratio of interfacial tensions and contact angles.

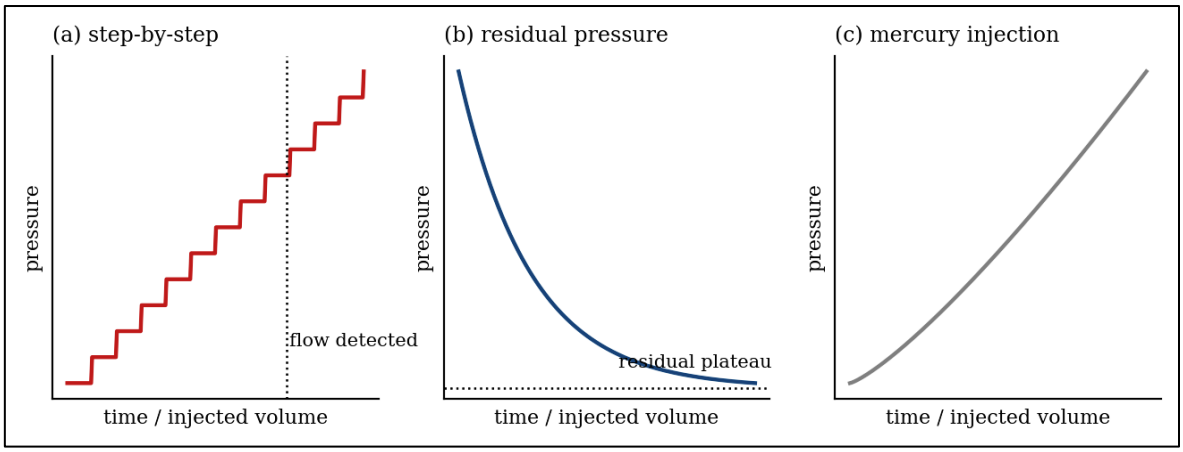


Figure 2.5 — Pressure signatures of the laboratory procedures.

Direct comparisons on the same samples give a consistent ordering: residual below dynamic below step-by-step. Egermann et al. (2006) and Boulin et al. (2013) each measured three or four procedures on single plugs; the drainage-side procedures agreed within tens of per cent, while the residual method gave about half the step-by-step value. Ito et al. (2011) reports the residual method giving 50 to 90 per cent of the step-by-step value. Vespo et al. (2024), simulating the procedures numerically, quantify the biases in both directions: the dynamic method overestimates by 13 to 208 per cent and the residual method underestimates by 6 to 70 per cent.

The mercury conversion is the least reliable step, and several studies report problems with it. Hildenbrand et al. (2002) measured both mercury-injection and gas displacement pressures on a suite of mudstones and clays and found no correlation between the two. They attribute the mismatch to the anisotropy of mudrocks, which omnidirectional mercury intrusion cannot capture, and to the drying at 105 °C the method requires, which alters clay fabric. Boulin et al. (2013) judged mercury injection unsuitable for an accurate entry pressure, noting that it is isobaric at zero effective stress while the entry pressure depends on the stress state. Espinoza & Santamarina (2017) note that mercury intrusion under-estimates in rocks with non-uniform pore-size distributions and cannot resolve bedding anisotropy. Klimkowski & Smulski (2012) add that pores below about 2.1 nanometers are invisible to the method and that a zero contact angle is unrealistic given the mixed wettability that Chiquet et al. (2007) document.

A second source of uncertainty in the mercury conversion is rarely stated. A mercury entry pressure is read from a curve, and the point on that curve taken to define it is a matter of convention. Kawaura et al. (2013) state theirs explicitly as the capillary pressure at ten per cent of pore volume occupied by mercury. Medina et al. (2019) report both a first-intrusion threshold and the pressure at twenty per cent saturation. Winland's index uses 35 per cent. Others use a tangent construction at the inflection point. These give materially different numbers from the same sample: in the Medina data the first-intrusion reading sits one to two orders of magnitude below the twenty-per-cent reading, because the first mercury to enter a plug fills surface irregularities and not the pore network. This convention effect can be of comparable size to the contact-angle effect, and Chapter 3 documents how it is handled.

Table 2.1 — Laboratory procedures for characterizing caprock sealing pressure.

Method	Curve interrogated	Quantity obtained	Relative value on the same sample	Role in this work
step-by-step	*drainage*	*breakthrough pressure*	*highest of the flow-through methods*	*source of the breakthrough (S1) target*
residual pressure	*imbibition*	*snap-off pressure*	*lowest; about 20 to 50 % of breakthrough*	*source of the snap-off (S2) target*
dynamic	*flow rates extrapolated to zero*	*extrapolated entry pressure*	*intermediate; overestimates if stabilized discharge is used*	*context only*
mercury injection	*intrusion*	*converted entry pressure*	*convention-dependent; poor agreement with gas methods*	*input to the quick-estimation chart (Chapter 4)*

These comparisons underpin two decisions. Breakthrough pressure and snap-off pressure are kept as separate targets and are never pooled, because they are separate physical quantities obtained by separate procedures. And mercury injection is kept out of model training, because its indirect entry pressure carries a different systematic bias from the flow-through methods. It is used instead in Chapter 4 as the basis of a quick-estimation chart.

2.9 The Thomas (1968) correlation and its limitations

The long-standing engineering approach to estimating breakthrough pressure without a dedicated capillary measurement is to correlate it with permeability through a power law. Thomas (1968) proposed a relation which Rocca (2021) records for nitrogen in field units as

$$P_{Th} = 7.37\, k^{-0.43} \qquad (2.5)$$

with pressure in psi and permeability in millidarcy. It was derived from five sandstone, two limestone and one dolomite plug, over a validity range of roughly 40 nanodarcy to 0.18 millidarcy and 0.06 to 5.17 MPa. Boulin et al. (2013) quote the same relation with an exponent of 0.45, a small discrepancy in the literature. A per-study fit shows the scatter the single form absorbs, though on a different quantity: Hildenbrand et al. (2004) fit their minimum capillary displacement pressure P_d (the residual pressure that persists at shut-off after breakthrough, a snap-off-type quantity rather than the drainage entry pressure, with P_d in pascals and k_abs in m²) and report log P_d = −0.72·log k_abs − 14.14 on their nitrogen experiments (coefficient of determination 0.50) and a slope of −0.41 on their carbon-dioxide experiments (0.40). The power-law form holds within a study; the coefficients differ between studies, between gases, and between the entry and residual pressures, which is why this work never pools the two.

The functional form has a physical rationale. Representing the rock as a bundle of capillary tubes and combining the Poiseuille permeability of the bundle with the Laplace entry pressure yields a

decreasing power law of permeability. This is why a permeability power law captures the first-order trend at all.

The same derivation exposes the correlation's limitations. The bundle-of-tubes model collapses the entire pore structure to a single effective radius. It ignores connectivity, tortuosity, pore-shape variation, the breadth of the pore-size distribution, and the fluid-specific interfacial tension and wettability that the Laplace equation shows to be first-order controls. Chen et al. (2024) state the assumption plainly, that the rock is treated as a porous medium composed of capillaries of the same diameter, and conclude that empirical models of this kind were established on specific samples from specific areas, so that heterogeneity of pore tortuosity and specific surface area limits their applicability.

Refitting the constants does not remove this limitation. A permeability-only correlation cannot represent the additional petrophysical, structural, and fluid-interaction controls that the physics identifies, and the evidence of Section 2.6 is that the relation can invert entirely on rocks whose microstructure differs.

Rocca (2021) is the closest precedent for this thesis. She collected and categorized literature threshold-pressure and residual-pressure data from seven studies, all nitrogen and brine at in-situ conditions, classified by laboratory procedure and by lithology, and tested the Thomas correlation against them. She reaches three conclusions. The two procedures agree acceptably within a permeability window of roughly 10^2 to 10^5 nanodarcy. Discrepancies are more case-sensitive than attributable to the procedure or the lithology. And Thomas is a reliable predictive approach for the identification, at least, of the order of magnitude of the threshold pressure. Section 2.17 states how the present work extends that analysis.

Part II — Methods for compiled data

2.10 Statistical methods for compiled datasets

Compiled petrophysical data is heterogeneous and rarely normally distributed. Permeability spans several orders of magnitude, distributions are skewed, outliers are common, and records originate from different rocks, methods and conditions. Rank-based and non-parametric approaches are therefore preferred throughout.

Correlation is measured two ways. The Pearson coefficient measures linear association and assumes approximately normal, outlier-free variables. The Spearman (1904) rank coefficient operates on the ranks, measuring monotonic association of any form, and is robust to outliers and to consistently directed non-linear relationships. It is the more suitable coefficient for a variable such as permeability.

Group differences are assessed with the Kruskal–Wallis (1952) test, the non-parametric equivalent of the analysis of variance. It ranks all observations jointly and tests whether mean rank differs between groups by more than sampling variation would produce.

Three further tools support Chapter 3. Outliers are flagged by the Tukey (1977) rule at 1.5 interquartile ranges, applied to the base-ten logarithm of each target. The dependence of a quantity on its source study is measured by decomposing its variance into between-study and within-study parts; a large between-study fraction indicates a quantity governed more by which study reported it than by the rock. And the concentration of a compilation across its sources is summarized by the effective number of sources, the inverse of the Simpson index.

A statistical test bounds only one kind of error. With the small, unevenly sized groups typical of compiled sealing data, a real effect of moderate size can go undetected, so a non-significant result is read here as an absence of evidence at the available sample size and weighed against the effect size and the physical expectation.

2.11 Machine-learning concepts

Machine learning learns a predictive relationship from data instead of from a functional form fixed in advance. Breakthrough-pressure prediction is a supervised regression problem: every training example carries a known continuous target, and the method learns to map inputs to it.

A supervised model is fitted to part of the data and judged on data it has not seen. Where the dataset is too small to give up a fixed validation partition, cross-validation is used: the data is divided into folds, and the model is trained and evaluated repeatedly with a different fold held out each time.

The central difficulty on a small dataset is the balance between under-fitting and over-fitting. A model that is too simple fails to capture the relationship. A model that is too flexible fits the noise as well as the signal, so it performs well on the training set and badly on anything new. The second failure is the harder to detect, because it hides behind high training accuracy, and the risk grows as the number of free parameters approaches the number of training examples.

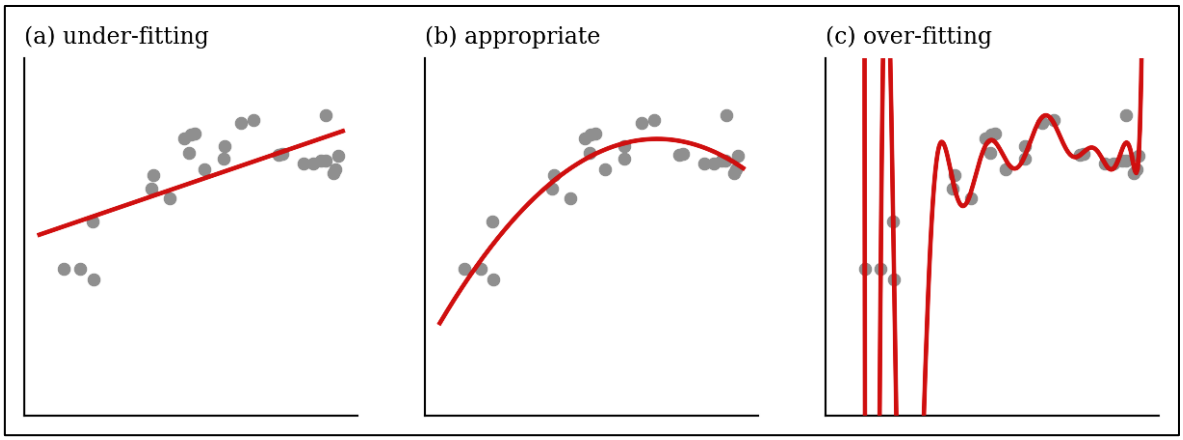


Figure 2.6 — Under-fitting, appropriate fitting, and over-fitting of a model to noisy observations. The third panel passes through every point including its noise.

Accuracy is reported as the coefficient of determination, R^2, the proportion of the variance in the target that the model reproduces. A value of 1 is perfect prediction and 0 is a model no better than always predicting the mean; negative values arise when a model performs worse than that baseline

on unseen data. Chapter 6 shows this measure has a specific weakness here, because it depends on how much variance the evaluated population contains.

Two reviews set out how these concepts apply in geoscience. Bergen et al. (2019) note that labels are often subjective or biased toward well-characterized phenomena, that interpretability matters more here than in image recognition, and that biased data, mixing training and testing data, overfitting and improper validation all lead to unreliable results. Xu et al. (2022), writing from industrial petrophysics, add that a sophisticated model applied to a small petrophysical dataset can produce seemingly good but erroneous results, and that machine learning may show no large advantage over conventional physical models in accuracy.

2.12 Data-driven prediction of breakthrough pressure

Machine learning has been applied to breakthrough-pressure prediction in a small number of studies. Hao et al. (2021) compiled 158 mudstone and shale data sets plus 14 sandstone data sets from sixteen publications (the thirteen rows of its Table 1, three of which bundle two papers each), converted values measured with different gases and with mercury to a common carbon-dioxide and brine basis using interfacial-tension ratios, and screened the correlation of breakthrough pressure with seven properties. Breakthrough pressure was most closely related to permeability and maximum throat radius, then to porosity and specific surface area, and essentially unrelated to density, depth, and organic carbon. A back-propagation neural network trained on four properties reached coefficients of determination of 0.992 and 0.989 on training and test sets.

The reported accuracy carries a qualification. The test set of fifteen samples was drawn at random from the same compiled pool as the training data, which came from thirteen different studies. A random split places records from the same study on both sides of the partition. The figure therefore measures interpolation among samples similar to those the model was trained on. It does not measure transfer to an unseen source. The contrast within the same paper makes the point: their pooled permeability correlation reaches 0.49 while their random-split network reaches 0.99 on the same data. The same study found that porosity's correlation, strong within a single source, nearly vanished once several sources were pooled. That is itself an argument that a random split distributes correlated same-source samples across both sides.

Wu et al. (2026) predicted breakthrough pressure for a shale caprock directly from well logs, taking gamma ray, acoustic transit time, bulk density and neutron porosity as inputs and core-measured breakthrough pressure as the target. Five algorithms were compared against a two-variable gamma-ray and acoustic regression baseline, on a held-out 30 % sample split and on four wells kept out of training altogether. AdaBoost performed best, reaching 0.89 on the held-out samples and 0.94 on the external wells, against 0.71 and 0.54 for the regression baseline. The external-well test is a genuine test of transfer and is the study's principal strength. Its scope is narrow. All four wells lie in the same depression and the same stratigraphic unit, so they differ in location but not in rock type, measurement method or source study, and the authors describe the result as regional external validation rather than evidence of cross-basin transferability. They also state that the 30 % split was

drawn at sample level and not by well, so core plugs from neighbouring depths in one well can fall on both sides of it, and they recommend leave-one-well-out validation as future work. That is the same concern, at well scale, that motivates the source-aware scheme adopted here.

Karami et al. (2025) reformulated the classical parachor model for gas–water interfacial tension using machine learning, fitting each fluid separately on a few tens to a hundred points and reaching per-fluid coefficients of determination up to about 0.99. The target is interfacial tension and not breakthrough pressure, so it is a methodological parallel and not a direct precedent.

Table 2.2 — Data-driven studies of breakthrough pressure and related targets.

Study	Target	Method	Validation	Limitation
Hao et al. (2021)	*breakthrough pressure, common basis*	*back-propagation neural network*	*random split within a pool of 16 publications (0.99 train, 0.99 test)*	*test set drawn from the same pooled data; measures interpolation, not transfer*
Karami et al. (2025)	*gas–water interfacial tension*	*ML reformulation of the parachor model, per fluid*	*held-out split, 70/15/15 and 80/10/10 (up to 0.99)*	*target is interfacial tension; a methodological parallel*
Wu et al. (2026)	*breakthrough pressure from well logs*	*five algorithms, AdaBoost best; GR, AC, DEN and CNL inputs*	*held-out sample split and four external wells (0.89 / 0.94)*	*one depression and one unit; the split is made at sample level, not by well*

Chen et al. (2024), reviewing the field as a whole, reach the same conclusion about empirical and data-driven models alike. They were established on specific samples from specific areas, and heterogeneity limits their applicability elsewhere.

2.13 Regression methods for small tabular data

Several method families are available for a small-sample tabular regression, each with a different trade-off between flexibility and data demand.

Linear and log-linear regression, the basis of the Thomas correlation, needs the fewest data points and stays fully interpretable, but it can only represent the form written into it in advance. Artificial neural networks can approximate an arbitrary non-linear function given enough data; on tens of rows that same flexibility becomes a liability. Support vector regression and Gaussian process regression are established choices for small tabular data and both handle non-linearity, but neither offers the advantage tree ensembles have on imperfectly scaled petrophysical inputs: a tree split uses only rank order, so the features need no common scale and an outlier in one does not distort splits made on another.

Tree ensembles occupy a middle ground, more flexible than a linear model and less data-hungry than a network, and interpretable through feature-importance analysis. Xu et al. (2022) add a

caution that bears on the choice: better results can sometimes come from a small but highly relevant training set than from a larger pooled one, and a model trained on one field or rock type usually applies only to that field or rock type. Chapter 5 tests whether flexibility helps here, and finds that it does not.

2.14 Random forests and gradient boosting

A random forest (Breiman, 2001) fits an ensemble of regression trees, each grown on a bootstrap resample of the rows and each choosing splits from a random subset of features. Because the trees are decorrelated, averaging them reduces the variance that makes any single deep tree unstable, and the forest can represent interactions and thresholds without those being specified in advance.

A gradient-boosted regressor (Friedman, 2001) grows trees in sequence, each fitted to the residual errors left by the ensemble so far and added at a small weight. Boosting typically extracts more signal from small datasets than bagging, and for the same reason it will fit noise if allowed to run too deep or too long. Shallow trees, a low learning rate and a limited number of rounds are the standard safeguards.

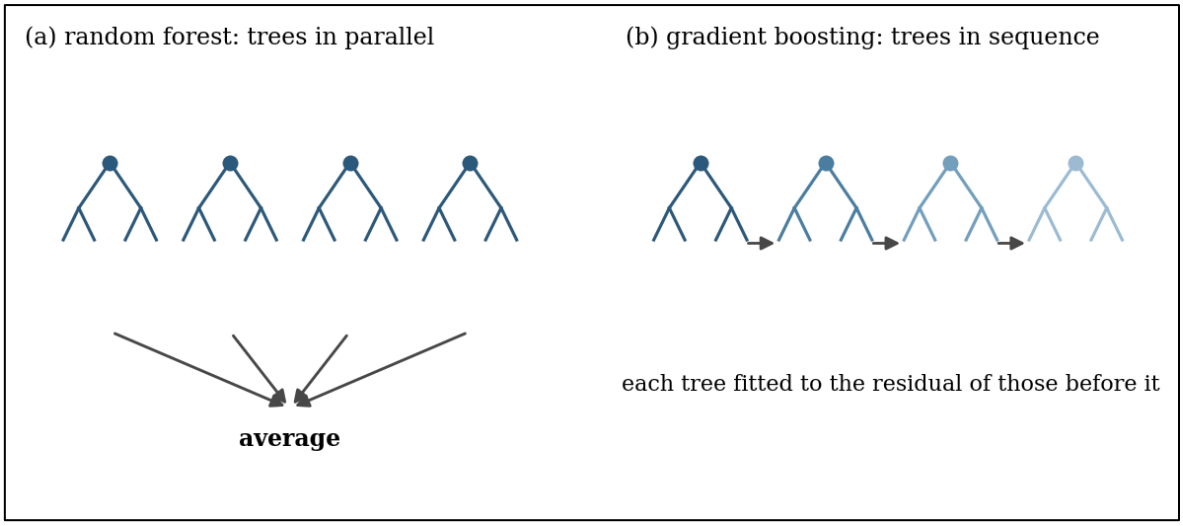


Figure 2.7 — Two tree-ensemble strategies. A random forest trains independent trees in parallel and averages them; gradient boosting adds trees in sequence, each fitted to the residual error of those before it.

Both methods have their settings fixed here by physical reasoning and sample size rather than tuned. Xu et al. (2022) note that petrophysical datasets of this size and heterogeneity are easily over-fitted by flexible learners, and searching hyperparameters against the same metric later reported as the headline result would bias that estimate. There is no separate tuning partition to search against without further shrinking an already small sample, so the configurations are reasonable defaults rather than optimized ones.

2.15 k-means clustering and cluster validation

Where no lithology label captures the sealing-relevant structure, an unsupervised classification can group samples by their own petrophysical similarity. K-means (MacQueen, 1967) partitions samples into k clusters by iteratively assigning each point to the nearest centroid and recomputing each centroid as the mean of its assigned points. Applied to porosity and log-permeability it groups rocks by their position in that plane instead of by a collapsed index, so a clay and an evaporite of comparable flow quality need not fall together. That is the limitation of the Winland-type indices of Section 2.7 which direct clustering avoids.

The number of clusters must be fixed in advance, and the silhouette score (Rousseeuw, 1987) is the standard diagnostic. For each point it compares the average distance to points in its own cluster against the average distance to points in the nearest other cluster, scaled to lie between −1 and 1. Where an independent reference grouping exists, the Adjusted Rand Index (Hubert & Arabie, 1985) scores agreement after correcting for chance. Section 3.11 shows that on a dataset of this size neither statistic settles the choice on its own.

2.16 Source-aware validation

The way a model's accuracy is estimated determines whether a reported score reflects genuine predictive ability or merely the reuse of familiar data. This is the methodological core of the thesis.

The pooling effect noted in Section 2.6 changes how accuracy should be estimated, and not only how inputs should be chosen. Take a correlation that is genuine within a single study but weakens once several studies are pooled. A conventional random train and test split places records from the same study on both sides of the partition. A model can then score well simply by recognizing the study's specific characteristics, its sample preparation, its particular rock type, its narrow range of confining pressure, without learning the underlying sealing physics at all.

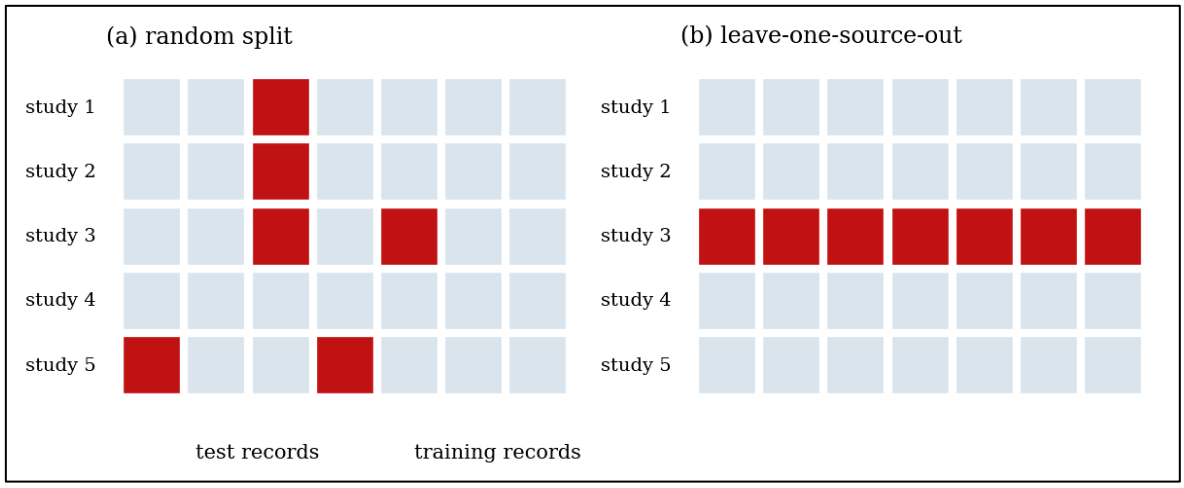


Figure 2.8 — Random versus source-aware partitioning of a multi-source dataset.

Xu et al. (2022) demonstrate the underlying effect directly on petrophysical data. They show a neutron-porosity and bulk-density correlation degrading as wells are added, from one well to three

to six. Data quality deteriorates as records accumulate from different wells and fields, because they are not measured under the same protocol and the geological conditions differ. Bergen et al. (2019) name mixing training and testing data among the practices that lead to unreliable results in geoscience machine learning.

The appropriate estimate for this data structure is leave-one-source-out cross-validation, in which each contributing study is withheld in turn and predicted from a model trained on all the others. Every prediction is then made on rocks from a study the model has never encountered, which is the same situation as applying a model calibrated on published data to a new, unmeasured site. The gap between a random-split score and a leave-one-source-out score is informative. It indicates how much a model depends on recognizing its training sources, and it is reported alongside conventional accuracy in Chapter 5.

Source-aware validation depends on the source attribution being correct. A record labelled with the wrong study places that study's data on both sides of the split and inflates the estimate, which is the failure the design exists to prevent. This is the practical reason the verification of Chapter 3 was undertaken, and Chapter 3 documents one block of twenty-five records where exactly that condition held.

Bergen et al. (2019) identify a further consequence for the field. The absence of ground truth and standard benchmarks means each new method is applied to a different dataset with author-defined criteria, which makes it difficult to determine relative performance and prevents the community from iterating on the best approaches. Chapter 6 returns to this in its recommendations.

2.17 Research gap

The physical picture of Sections 2.1 to 2.9 and the modeling toolkit of Sections 2.10 to 2.16 can now be drawn together.

The properties that govern breakthrough pressure fall into four categories: matrix properties describing the rock, external conditions describing the state it is held in, fluid properties describing the two phases, and fluid–rock interaction properties, principally the contact angle, which depends on both mineralogy and fluid chemistry and is a property of neither alone.

Table 2.3 — Property categories governing breakthrough pressure, and their use.

category	***properties***	***Thomas (1968)***	***this work***
matrix	*porosity, permeability, pore-throat radius, specific surface area*	*permeability only*	*permeability and porosity*
external	*confining stress, temperature, water saturation*	*not used*	*examined and set aside (Ch. 3, 5)*
fluid	*interfacial tension, phase state*	*not used*	*handled as a conversion (Ch. 4)*
fluid–rock	*contact angle*	*not used*	*not available at sample level*

Judged against this framework, the established empirical tool uses only a small part of it. The Thomas correlation relates breakthrough pressure to a single matrix property through a power law whose physical derivation is exactly why it captures a first-order trend and exactly why it cannot capture more. Matrix properties beyond permeability, principally porosity, carry information a permeability-only correlation is structurally unable to use.

The first is methodological. Machine learning offers a way to combine multiple properties without a functional form fixed in advance, and the precedents of Section 2.12 show that data-driven approaches can extract a usable signal. What those precedents leave unresolved is whether that signal survives contact with a rock type, study, or measurement method absent from training. Every published figure for this problem rests on a random split of pooled data, or, in the one study that held wells back, on wells from a single depression and a single stratigraphic unit.

The second is evidential. Rocca (2021) is the closest precedent, and it establishes that the Thomas correlation is reliable to the order of magnitude on nitrogen data assessed by visual agreement. It does not quantify how far the relation transfers, does not separate breakthrough from snap-off as distinct targets, and covers one fluid. The present work extends it in each of those directions.

The third part of the gap is one the thesis did not set out to address but could not avoid. A compiled dataset carries the errors of its compilation, and those errors are not visible in summary statistics. Chapter 3 documents values entered in the wrong column, permeabilities derived and not measured directly and wrong by three orders of magnitude, records whose cited source contains no such measurement, and a block whose experimental conditions were placeholder values. Nearly two thirds of the merged compilation did not survive verification. Any statement about how far a model transfers is conditional on the records being what they claim to be, and establishing that is prerequisite to the modelling and not preliminary to it.

A further limit is physical rather than methodological. Breakthrough pressure and snap-off pressure are governed by related but distinct pore-scale properties, respectively the narrowest throat on the widest connected path and the pore-body-to-throat aspect ratio. Only the first is well constrained by porosity and permeability. The two are therefore distinct problems, and neither quantity can be inferred from the other.

Chapter 3 takes up the dataset: how records assembled from many studies were compiled, verified against their primary sources, screened, and analysed against the properties and methods introduced in this chapter.

Chapter 3 — Compilation, Verification and Analysis of the Dataset

The models developed in the following chapters are built on the dataset verified here. The dataset was provided by the supervisor as a compilation of caprock sealing-pressure measurements drawn from the experimental literature and organized by fluid system. This chapter describes how that compilation was consolidated, traced record by record to its primary publications and checked value by value, what was added to it, and how the records that survived screening behave. Part I covers the verification, the enrichment and the screening. Part II analyses the retained data: the two target pressures, their behavior across fluids and rock types, the relationships between features and targets, and a data-driven grouping of the samples. Every quantity is reported both as received and after screening, so that the effect of each decision is visible.

Figures follow one convention throughout: fluids are identified by color and marker (CO_2 circle, N_2 square, CH_4 triangle, H_2 diamond, mercury thin diamond); breakthrough pressure is red and snap-off pressure blue; the three seal-quality groups defined later in the chapter are green, amber and red in order of sealing capacity.

Part I — Verification, enrichment and screening

3.1 The compilation as received

The compilation was provided as a workbook of four sheets, one per fluid system, each listing sample-measurements with the test type, laboratory conditions, lithology, porosity, permeability, the interfacial tension and contact angle used, the sealing pressure obtained, and a reference. The four sheets were consolidated into one analysis table of 320 sample-measurements under a common vocabulary and unit system, with permeability in millidarcy and pressure in bar. Permeability spans seven orders of magnitude, so it is carried as its base-ten logarithm; on a linear scale a few high values dominate the distribution and obscure the rest.

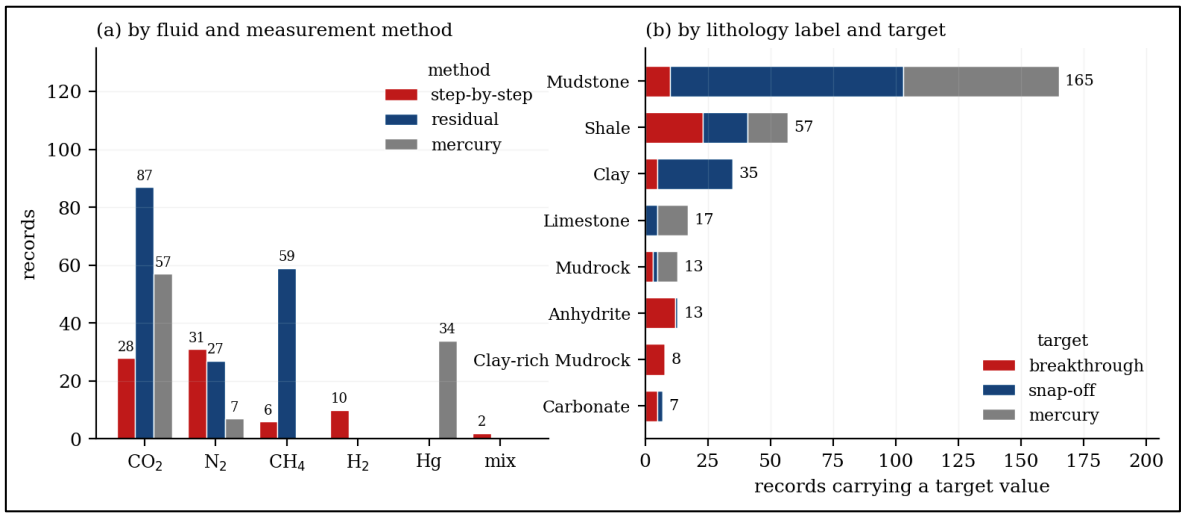


Figure 3.1 — The merged compilation as received: by fluid and measurement method, and by lithology label and target.

The compilation carries, for most records, the interfacial tension and contact angle that the source study applied. These are study-level constants and not sample measurements. Nitrogen carries 0° in every record and 72 mN/m in most. Mercury carries 140° and 480 to 485 mN/m throughout. Carbon dioxide carries 60° and 60 mN/m in the great majority, with 27 mN/m and 34° in the studies that worked at high pressure. They are literature defaults adopted by each study, and they are treated here as documentation of what each study assumed and not as measured properties.

3.2 Verification against primary sources

To establish rigorous data provenance, every record in the consolidated compilation was verified against its primary publication across 47 retrieved sources. Each entry was audited across four core dimensions; reported sealing pressure and units, measurement scale and support volume, experimental context (including fluid system and applied stress), and source study identity. All reported pressure units (bar, kPa, MPa, and psi) were standardized to bar and cross-checked against original published figures to eliminate conversion errors.

This record-by-record audit identified errors across ten source blocks, which were rectified against published tables. Corrective actions included updating miscited references, resolving misreported pressure ranges and confining stresses, eliminating duplicate entries, correcting mis-transcribed porosities, replacing mercury-derived permeabilities with direct core measurements, and rebuilding fragmented plug identifiers. Additionally, four records from Gao et al. (2014) were reclassified from breakthrough pressure to snap-off pressure based on their pulse-decay residual testing methodology. Finally, ten record blocks failing basic traceability or data-quality criteria were identified and excluded from predictive modeling.

3.3 Enrichment of the compilation

Five further studies were identified and added, each verified against its publication to the same standard. Together they extend the compilation in three directions the original lacked: hydrogen, a wider stress range, and controlled intact-against-altered comparisons. Not all five reach the modelling populations. The Kim, Vilarrasa & Makhnenko (2025) values are read from the published figures and not from a table, and they are excluded at the screening of Section 3.4 under the figure-derived ground (Table A.3); the Benlalam et al. (2026) records are excluded for the reason given at the end of this section. Three of the five added studies therefore enter the modelling populations.

Table 3.1 — Studies added to the compilation.

study	records	fluids	what it adds
Li, Gates & Hejazi (2025)	*9*	*N_2, CH_4, H_2*	*hydrogen breakthrough on brine-saturated plugs with measured porosity and permeability*
Rezaeyan et al. (2015)	*9*	*CO_2, N_2, CH_4*	*shale and anhydrite at 241 to 379 bar*
Benlalam et al. (2026)	*8*	*H_2, CH_4*	*hydrogen on a clay-rich caprock*

study	*records*	*fluids*	*what it adds*
Makhnenko et al. (2017)	*6*	*CO_2*	*intact and remoulded Opalinus Clay*
Kim, Vilarrasa & Makhnenko (2025)	*6*	*CO_2*	*intact, fractured and remoulded facies*

The five studies listed sum to 38 records. The remaining 55 of the 93 added records are the literature compilation of Hao et al., added during enrichment and removed as a block at the screening of Section 3.4 because it is a secondary source.

With the value corrections and restored measurements described in Section 3.2 and Appendix A applied, the merged table holds 413 records: 320 from the compilation and its sources and 93 added.

Two choices about the Li, Gates & Hejazi (2025) records need stating. Their paper reports both a breakthrough pressure and a threshold pressure, the latter 14 to 25 per cent lower across their plugs; the breakthrough-pressure column is the one used. And their confining pressure was held 5 MPa above the injection pressure throughout the test (Li et al., 2025, §2.2), so the confining pressure recorded at breakthrough equals the breakthrough pressure plus 50 bar. It is retained in the dataset as a real applied stress, but it is flagged and never used as a model input, because it encodes the target.

The compilation as originally received contained no hydrogen measurement, and Chapter 1 identifies that absence as a limitation on extending any result to hydrogen storage. Two of the added studies remove it. Benlalam et al. (2026), a study co-authored by the thesis supervisor and by the storage operator Snam-Stogit and cited here with that interest disclosed, report nine step-by-step threshold-pressure experiments with hydrogen on the Argille del Santerno Formation, seven of which are compiled here with one of their methane measurements, and Li, Gates & Hejazi (2025) report hydrogen alongside nitrogen and methane on the same outcrop shales. The two studies do not give the same answer. On the three plugs of Li, Gates & Hejazi (2025), each tested with all three gases, hydrogen broke through at a pressure 14 to 19 per cent below the nitrogen value on the same plug (ratios of 0.81, 0.86 and 0.81 for plugs E, M and W), while methane broke through within 5 per cent of nitrogen. Benlalam et al. (2026) report hydrogen and methane threshold pressures on their caprock as comparable. The hydrogen result carried forward from these data is therefore that hydrogen seals at a pressure close to, and on the one same-plug comparison somewhat below, that of nitrogen and methane.

Not every added record is usable for modelling. The Benlalam records report permeability only as upper bounds and do not report porosity, so their petrophysical features would have to be assumed. They are retained in the dataset and cited for the hydrogen result, but they are excluded from the modelling populations under the same rule that excludes imputed features elsewhere in this chapter.

3.4 Screening criteria and their effect

The screening is applied to the merged compilation, and every exclusion rests on one of three grounds: physics, data quality or traceability, or a result of the analysis itself. Figure 3.2 gives the funnel and Table 3.2 the argument for each step.

Table 3.2 — Screening of the merged compilation.

step	removed	remaining	ground
Records as received	—	413	
Duplicate records	65	348	data quality
Compiled sources without primary petrophysics	134	214	traceability
Mixed-source block with placeholder conditions	25	189	data quality
Records without a traceable source	24	165	traceability
Values read from figures, not tabulated	6	159	traceability
Records without measured petrophysics	8	151	data quality
Crack-dominated or post-treatment measurements	3	148	physics
Reservoir rock outside caprock scope	2	146	physics
Verified foundation	—	146	

Compiled sources without primary petrophysics (134 records). Two blocks fall here. The larger is a review article tabulating basin-scale sealing estimates for Chinese gas fields: a formation-average sealing pressure over a field-scale support volume is a different physical quantity from a centimeter-scale plug measurement, and these records carry no porosity, permeability or depth. The smaller is a literature compilation added during enrichment. Reading it showed that it compiles sixteen publications (thirteen rows of its Table 1, three bundling two papers each), most of which are already primary sources in the present dataset; merging it would duplicate existing records under a second label, and its values are converted to a common basis and not reported as measured.

Mixed-source block with placeholder conditions (25 records). Cell-checking showed that the target values in this block are traceable to at least three different studies, but the features and conditions match none of them: for one subset the permeability is a factor of a thousand from the published value, the temperature is 31 °C against a published 90 °C and the confining pressure 78 bar against 241 bar. Every record in the block carries the same temperature and confining pressure, and those values are the critical point of carbon dioxide. Three records also duplicated measurements already present from other sources.

Records without a traceable source (24 records). Two are attributed to a publication that, on reading, contains no laboratory measurement at all. Six had been attributed to Kawaura et al. (2013) on the resemblance of their petrophysics to that study; on reading, the study reports one plug, and none of the six sealing pressures, porosities or permeabilities appears in it, so the attribution cannot be sustained and the records have no identifiable source. The remainder carry complete petrophysics

but no identifiable source. That is not only a provenance problem: the validation strategy of Chapter 5 groups records by originating study, and a block that cannot be attributed cannot be assigned a grouping key. One such block spans three lithologies and depths from 600 to 4000 m, so it is unlikely to be one study, and treating it as one group would be an assumption without basis.

Crack-dominated measurements (3 records). The source study states the reason itself: the measured pressures correspond to pore radii of four to six micrometres, far larger than the nanometre radii expected for those shales, and microscopy revealed hairline cracks parallel to bedding in every plug, probably formed during coring. One value is a post-treatment shut-off after slurry injection, not a rock property.

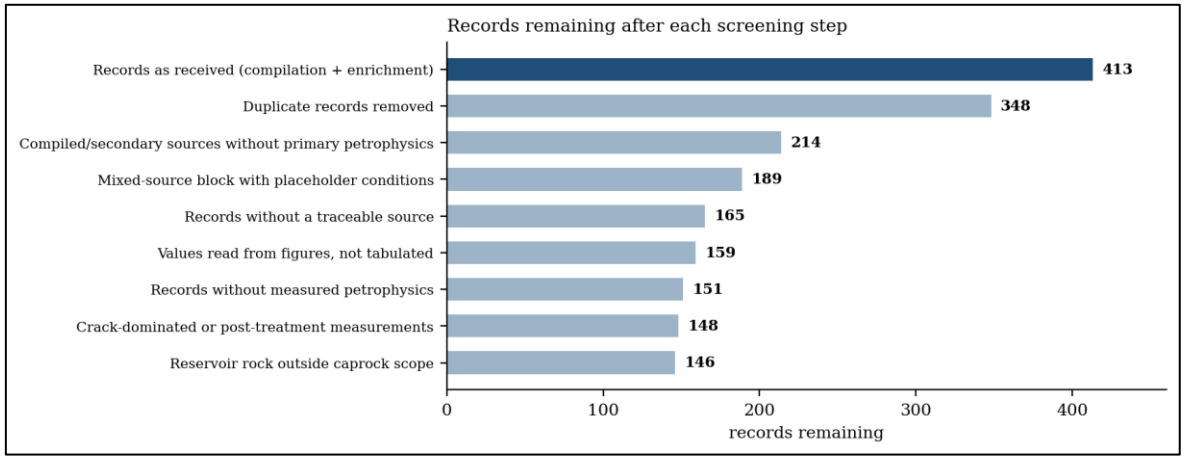


Figure 3.2 — Records remaining after each screening step.

Figure 3.3 puts the same result as proportions. Of the 348 non-duplicate records, 146 are retained. The single largest exclusion is the 134 records from compiled sources, which is 39 per cent of the whole. The verified foundation is dominated by nitrogen and carbon dioxide, at 48 and 39 records, with mercury as the fluid in 33, methane in 23 and hydrogen in 3. Forty records were measured by mercury injection; seven of them were converted to a gas basis by their authors and carry that gas as the fluid label.

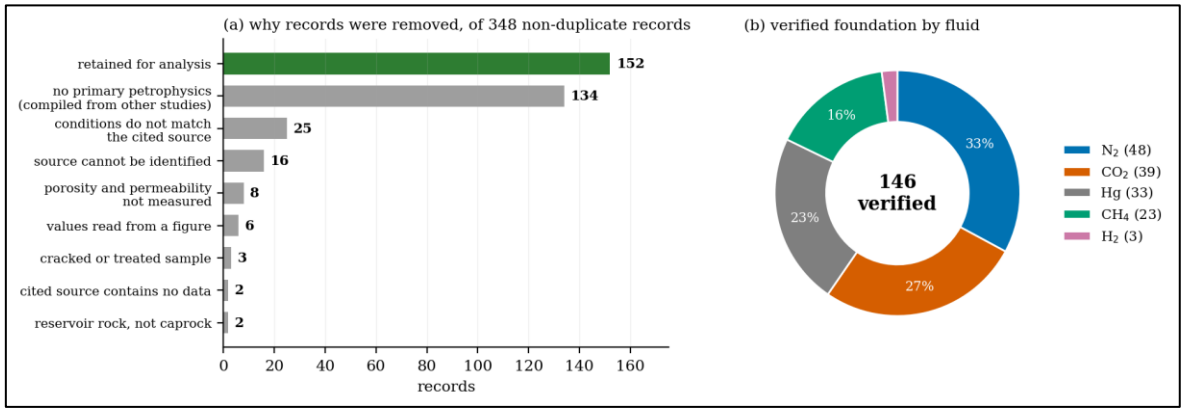


Figure 3.3 — Screening outcome and fluid composition of the verified foundation.

3.5 The verified foundation

The 146 verified records come from 14 source publications and represent 98 distinct plugs. They divide into three populations: 38 flow-through breakthrough measurements, 69 snap-off measurements and 40 mercury-injection measurements. The three sum to 147, not 146, because one record carries a step-by-step and a residual value on the same plug. Table 3.3 gives the distribution of each target.

Table 3.3 — Descriptive statistics of the two targets and the two features.

quantity	n	min	Q1	median	Q3	max	mean	SD	skew	CV
breakthrough pressure (bar)	38	0.90	10.8	41.5	189.9	343.5	98.4	115.7	+1.00	1.18
snap-off pressure (bar)	69	0.20	3.70	8.00	17.6	157.0	17.9	27.2	+3.10	1.51
mercury entry pressure, mercury–air basis (bar)	33	6.41	22.1	69.9	365.4	796.5	222.5	247.7	–	1.11
mercury entry pressure, gas basis as reported (bar)	7	4.00	5.9	13.0	29.4	102.3	27.1	36.1	–	1.33
porosity, S1 (–)	38	0.001	0.003	0.030	0.141	0.400	0.084	0.112	+1.27	1.34
$\log_{10}$ permeability, S1 (mD)	38	−5.18	−4.59	−3.04	−2.25	+0.24	−3.18	1.35	+0.24	–
porosity, S2 (–)	48	0.006	0.068	0.150	0.241	0.400	0.159	0.101	+0.08	0.63
$\log_{10}$ permeability, S2 (mD)	48	−6.52	−5.04	−3.91	−2.23	+0.31	−3.60	1.75	+0.49	–

Both targets are right-skewed, the snap-off distribution severely so at +3.10, and both carry a coefficient of variation above one. Neither is close to normal, which is why the analysis that follows uses rank-based statistics and why the targets are modelled as logarithms. The two modelling populations differ in composition as well as size. The breakthrough population reaches lower porosities, a first quartile of 0.003 against 0.068 for snap-off, because it contains the tight anhydrites and evaporites that carry the highest sealing pressures.

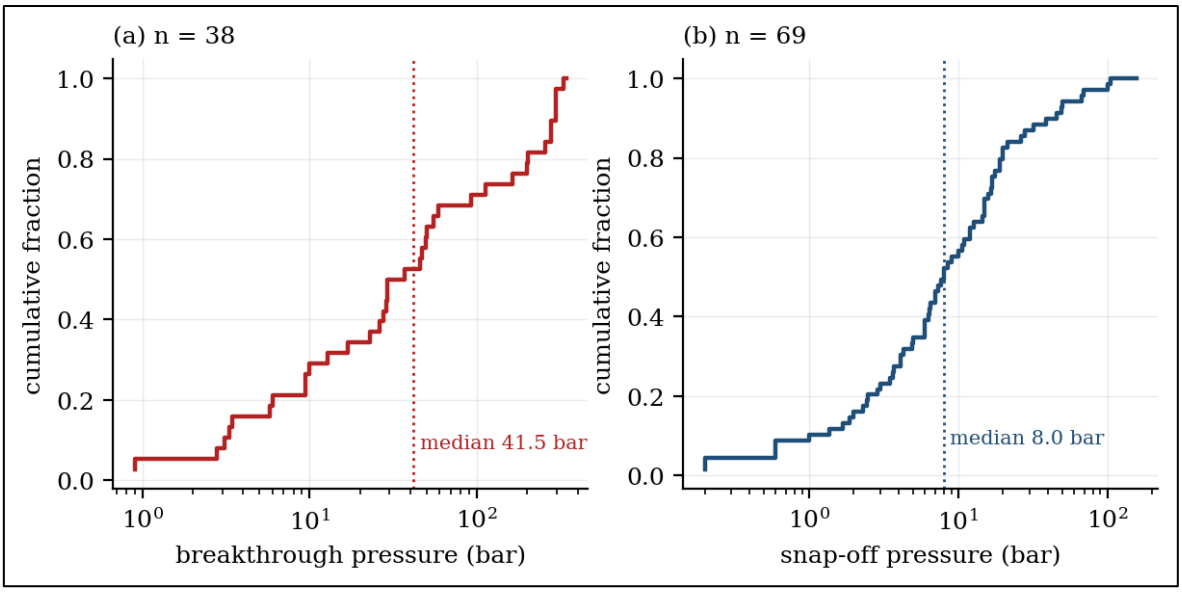


Figure 3.4 — Cumulative distributions of breakthrough and snap-off pressure.

The mercury population requires a note on which quantity it carries. The largest mercury source reports two pressures for each sample: a threshold pressure read at first detectable intrusion, and the pressure at twenty per cent mercury saturation. The two differ by a median factor of about sixty. Compared against every other mercury measurement in the dataset, which fall between 38 and 524 bar on a mercury-air basis, the threshold column, at 0.7 to 11.6 bar, sits one to two orders of magnitude low, while the twenty-per-cent value, at 6 to 796 bar, sits in the same range. The reason is conformance: the first mercury to enter a plug fills surface irregularities and not the pore network, which is why the standard practice is to read at a saturation point. One study in this dataset states its convention explicitly as the capillary pressure at ten per cent mercury saturation. The twenty-per-cent values are therefore adopted, reconstructed for all twenty-eight samples from the source table, and stored on the mercury-air basis so that they are converted by the same relation as every other mercury record.

The verified dataset is unevenly weighted across its sources (Figure 3.5). Fourteen groups of equal size would give an inverse-Simpson index of 14; the observed index is 7.9, and the three largest sources hold 75 of the 146 records. Any statistic pooled across the whole dataset is therefore weighted toward the rocks and conditions of a handful of contributors, which is why provenance is carried through the analysis and the models are validated source by source.

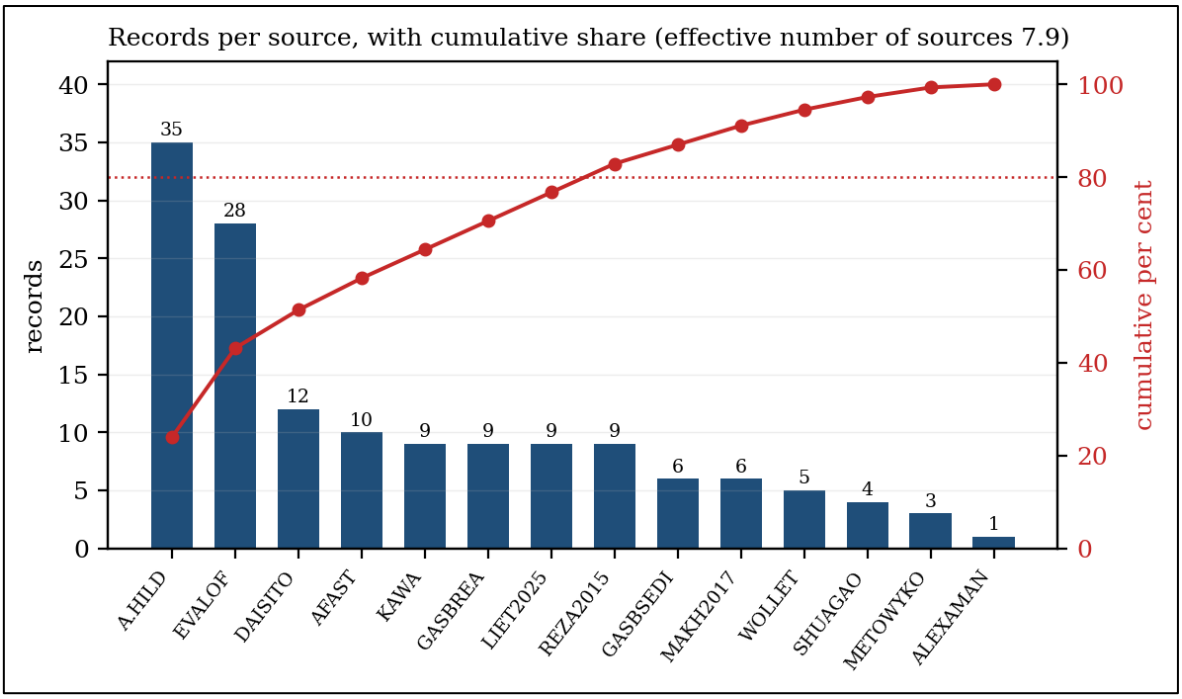


Figure 3.5 — Records per source, with the cumulative share.

Compiled petrophysical data are incomplete by construction. Each study reports the properties its own objectives required, and none reports every property a pooled analysis might use. Lithology is recorded for every verified record, porosity for 97 %, temperature for 93 %, permeability for 80 %, confining pressure for 77 %, and depth for 71 %.

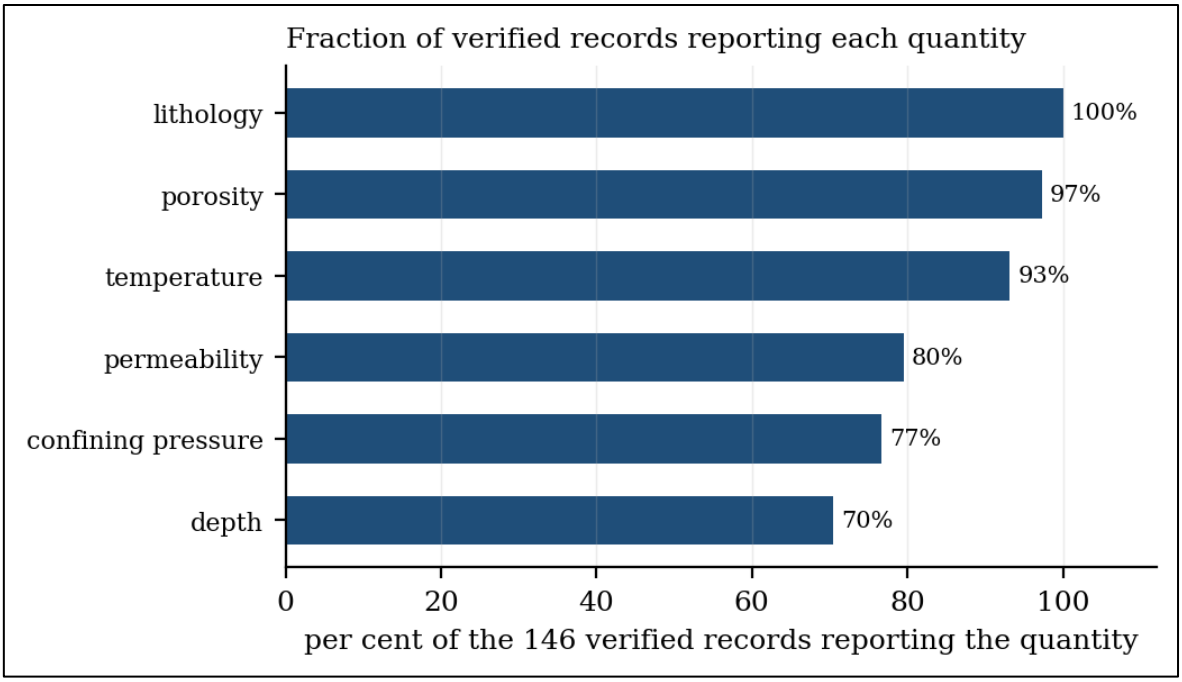


Figure 3.6 — Completeness of the verified foundation, by quantity.

The missingness is a property of the study and not of the sample. Porosity, temperature and depth are reported by each publication either for all of its samples or for none of them, and permeability and confining pressure in 86 % of publications. The gaps form horizontal bands, one study at a time.

Because missingness follows the study, the retained records are modelled complete-case, with no imputation of the target pressures and none of the input features. A study-level fill-in would give every sample in a study the same substituted value, and a model could then read that value as a marker of the study. The data show why depth cannot serve as a substitute: neither confining pressure nor temperature has a usable relationship with sample depth, because each study sets its own laboratory conditions independently of where the rock came from.

Of the forty mercury-injection records, thirty-three are reported on the mercury–air basis and seven were already placed on a gas basis by their authors. Only the thirty-three are converted, since applying the conversion to all forty would convert seven records twice, and Table 3.3 reports the two groups separately for the same reason.

Raw and converted values should not be pooled. As reported the two medians differ by a factor of 5.4, at 69.9 and 13.0 bar; read on one axis without conversion, the same records would suggest that mercury injection over-estimates entry pressure fivefold, which is a difference of basis and not a property of the method. The two groups are not independent of one another: the seven gas-basis values are the source authors' own conversions, by the same relation (Equation 4.2), of mercury readings from the same studies. Their agreement once both are on a gas basis is therefore arithmetic and not a test of the conversion. The only empirical tests of the conversion are the same-plug comparisons of Section 4.3.

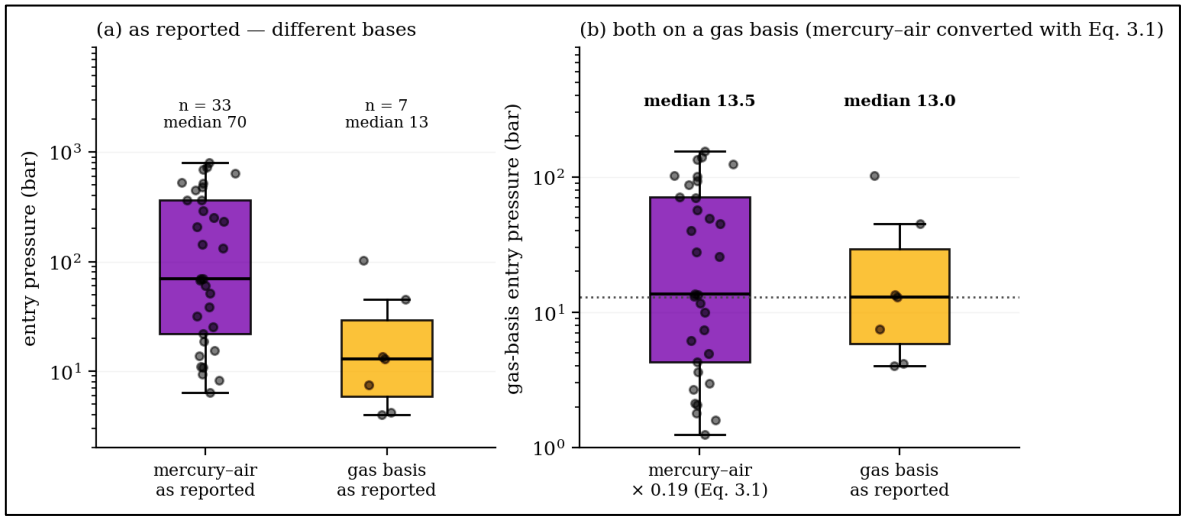


Figure 3.7 — Mercury entry pressures as reported and after conversion to a common basis.

Two attributes record distinctions the pipeline must respect. The first is the meaning of the confining-pressure column. For the mercury-injection block it holds the in-situ subsurface pressure of the formation the sample came from. Mercury injection is performed on a dry plug at ambient stress, so no confining pressure exists. For every other source the column holds the stress applied in the test. The second is the wetting phase: brine or water for every flow-through record, air for the mercury records. A study measuring threshold pressure against anhydrous kerosene was identified during the literature search and excluded on this ground, since a gas-kerosene threshold is a different quantity from a gas-brine one.

Part II — Analysis

3.6 Sealing pressures across fluid systems

Chapter 2 established on physical grounds that breakthrough and snap-off pressure are governed by different pore-scale mechanisms and are treated as separate quantities. This section tests whether the verified data behave as the physics requires.

The median snap-off pressure is 8.0 bar against a median breakthrough pressure of 41.5 bar, a pooled ratio of medians of 0.19, just below the 20 to 50 per cent range the literature reports. Section 3.8 shows why this pooled ratio is not a property of the rocks.

Fluid identity does not separate either target, and Table 3.4 gives the tests.

Table 3.4 — Kruskal–Wallis tests on the verified data.

comparison	groups	H	p	verdict at p < 0.05
breakthrough pressure by fluid	4	2.30	0.51	no difference detected
snap-off pressure by fluid	3	5.95	0.051	no difference detected
breakthrough pressure by lithology label	2	1.81	0.178	not separated
breakthrough pressure by seal-quality group	3	15.14	0.0005	strongly separated

A non-significant result at this sample size is an absence of evidence and not evidence of absence. The breakthrough population now contains four fluids, nitrogen, carbon dioxide, hydrogen and methane, and the spread within each fluid exceeds the difference between them by more than an order of magnitude. This does not contradict the ordering established on paired measurements, where nitrogen breakthrough is two and a half to three times the carbon dioxide value on the same rock. The fluid effect is real but it is a factor of about three, and it is visible only when two fluids are tested on the same sample. In a pooled compilation the fluids are measured on different rocks from different studies, and rock-to-rock variation spans seven orders of magnitude in permeability, so a threefold fluid effect is lost in the between-sample scatter. Fluid identity cannot be recovered as a feature from these data, and the interfacial tension and contact angle serve as conversion factors between fluids and not as model inputs.

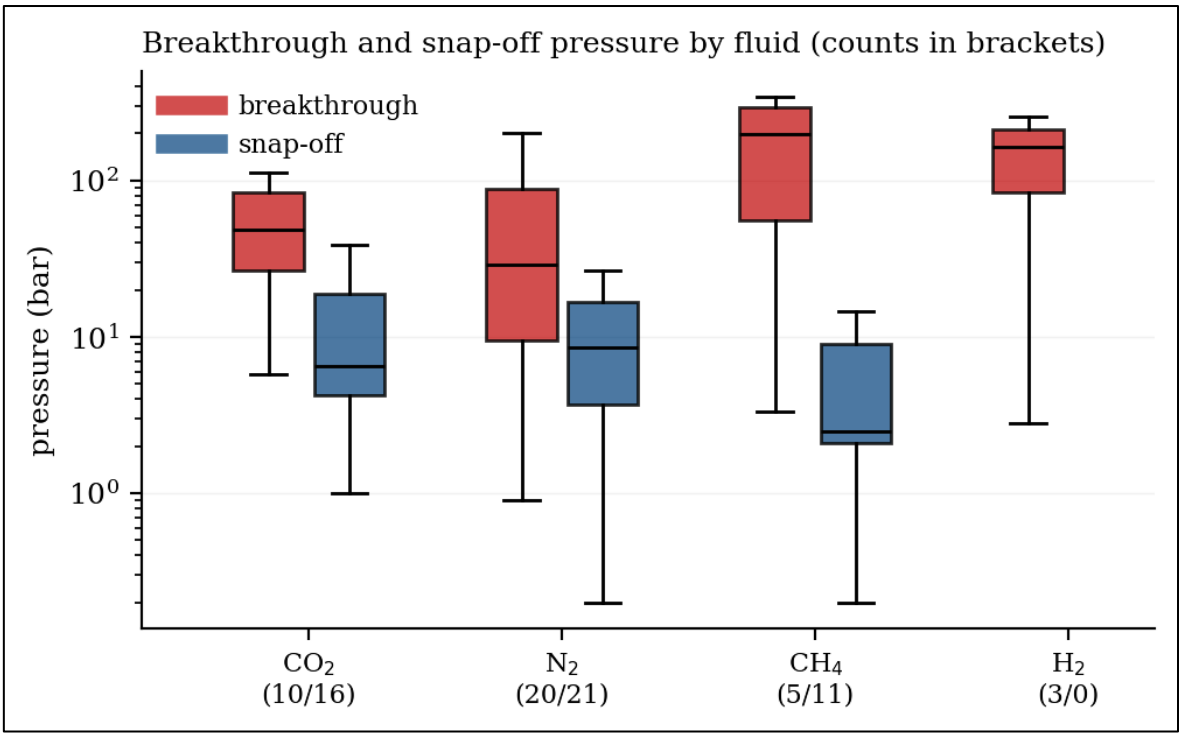


Figure 3.8 — Both pressures by fluid, with the ratio of medians.

Carbon dioxide records span both liquid and supercritical states, since the laboratory conditions sit above the critical pressure but on either side of the critical temperature (73.8 bar, 31.0 °C; Span & Wagner, 1996). Twenty-four records are liquid and fifteen supercritical. Phase controls the interfacial tension and through it the entry pressure, so the two populations are compared directly (Figure B.5). Their distributions overlap almost entirely, and the difference between the medians is small against the spread within either group. Phase is a real physical dimension of the data and it is not a dominant control across these samples, so it is not carried as a separate feature.

Grouped by lithology label the two pressures vary in the expected direction, but the medians are unreliable. Each label pools different fluids and methods in different proportions, and the counts behind several labels are small. The deeper problem is that one label covers very different rocks.

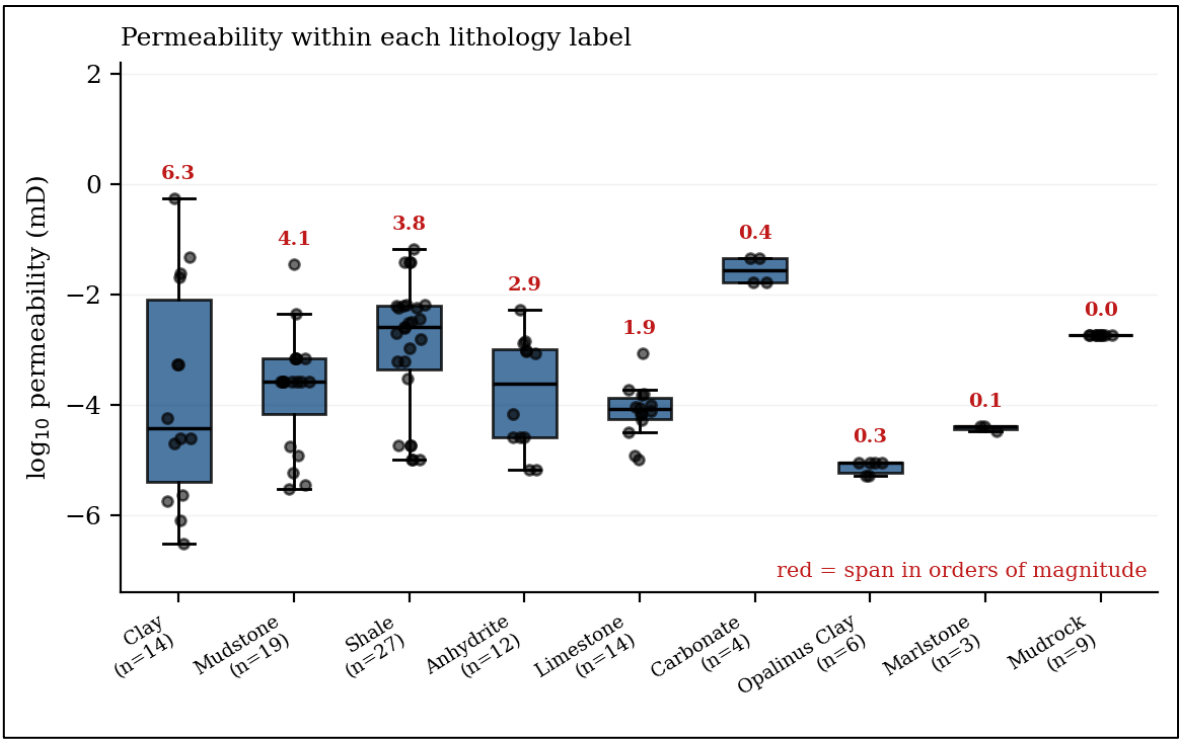


Figure 3.9 — Permeability span within each lithology label.

Permeability spans six orders of magnitude within the clay label, 4.1 within mudstone and 3.8 within shale. A named rock type does not fix the rock, and that is why Section 3.11 derives a grouping from the measurements themselves.

3.7 Feature–target relationships

Requiring complete records for permeability and porosity yields the two modelling populations: 38 breakthrough records from 8 publications, representing 29 plugs, and 48 snap-off records from 9 publications, representing 42 plugs. Every permeability in both populations is a direct measurement; the mercury-derived permeabilities corrected in Section 3.2 belong to a block that enters neither.

The features relate to the two targets in physically distinct ways, and the data follow the physics. Figure 3.10 compares the feature–target correlations of the two pressures by the Pearson coefficient, computed on the logarithm of pressure because the Thomas relation is a power law, and by the Spearman rank coefficient, which is insensitive to the several-order-of-magnitude spread of permeability.

Table 3.5 — Feature–target correlations.

target	feature	n	Pearson (log P)	Spearman	p (Spearman)
Breakthrough	$\log_{10} k$	38	−0.79	−0.73	< 0.0001
Breakthrough	porosity	38	−0.62	−0.76	< 0.0001
Snap-off	$\log_{10} k$	48	−0.35	−0.22	0.13
Snap-off	porosity	48	−0.03	+0.08	0.57

For breakthrough pressure both petrophysical features carry a strong and significant signal, and by rank correlation porosity is slightly the stronger of the two. The negative sign is the one the physics requires: permeability rises with the square of the controlling throat radius while capillary entry pressure varies with its inverse, so a rock of lower permeability has narrower throats and a higher entry pressure. Porosity correlates negatively for the same reason: within these sealing lithologies, lower porosity accompanies the finer pore structure that raises entry pressure.

This result bears directly on a contrast reported in the literature, where porosity correlated strongly within a single data source but almost vanished once data from many sources were pooled. In the present population porosity survives pooling at a strong and significant level. Two features of the population plausibly account for the difference: it excludes the mercury-injection records that a pooled compilation would otherwise mix with flow-through data, and every record has been traced to a primary source, so the population is not diluted by values of uncertain provenance.

The literature is not unanimous on the sign. Two studies report breakthrough pressure increasing with porosity. One explanation, not tested here, is that the sign of the porosity relation depends on what varies across the sample set. Where a suite spans a compaction and diagenesis trend, porosity and throat size fall together and the correlation is negative. Where it spans a lithofacies contrast, clay-rich rocks carry high porosity in micropores but very small throats, so porosity rises while throats shrink and the correlation reverses. The present population spans the first case.

For snap-off pressure the pattern differs. Permeability retains a weak association, significant on the Pearson coefficient but not on the rank coefficient at this sample size, while porosity shows no detectable association: −0.03 on Pearson and +0.08 on Spearman, both indistinguishable from zero. Porosity is informative for breakthrough pressure but not, on these data, for snap-off pressure.

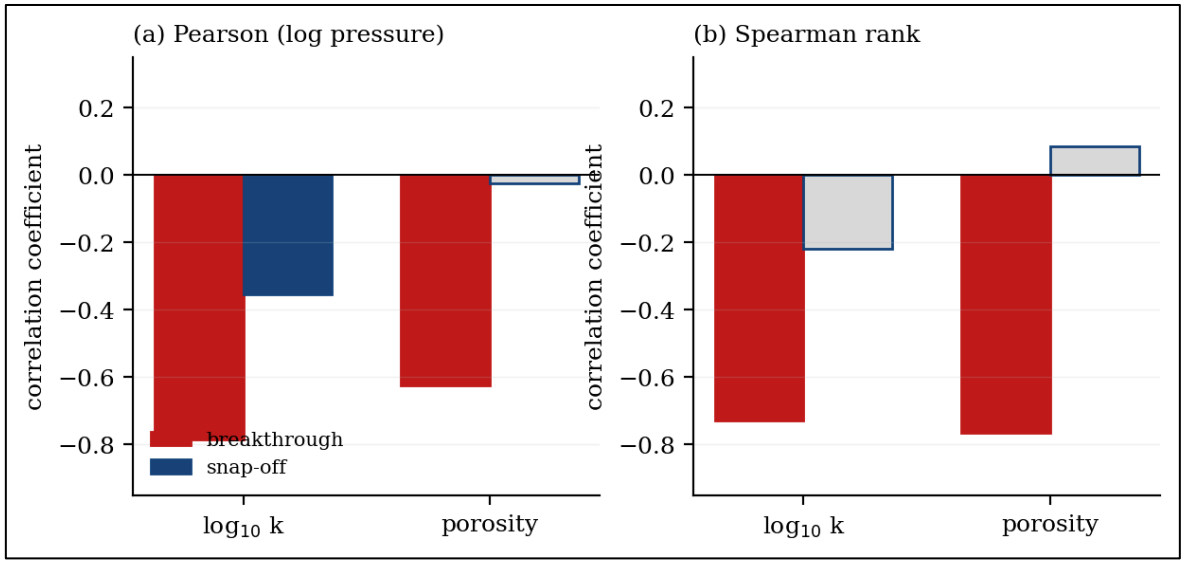


Figure 3.10 — Feature–target correlations for the two pressures.

3.8 Outliers and their treatment

Because the targets span orders of magnitude, outliers are assessed on the base-ten logarithm of each, with the Tukey rule flagging any value more than 1.5 interquartile ranges beyond the first or third quartile. The criterion depends only on rank order and is therefore robust to the skew documented in Table 3.3.

The choice to apply the rule to the logarithm matters as much as the rule itself, and Figure 3.11 shows it by running three criteria on the same data.

Table 3.6 — Records flagged by three outlier criteria.

criterion	breakthrough	snap-off	which records
Tukey, raw pressure	*0*	*4*	*the four highest snap-off values, 39 to 67 bar*
mean ± 2 SD, raw pressure	*1*	*3*	*the highest breakthrough value, 343.5 bar, and the three highest snap-off values*
Tukey, $\log_{10}$ (adopted)	*0*	*2*	*the two lowest snap-off values, both 0.2 bar*

The rules broadly agree on how many records are unusual but not on which ones. On snap-off, the two raw-scale criteria flag three or four records at the top of the distribution; the log-scale criterion flags two at the bottom. Not one record is common to both answers.

The reason is the shape of the data. Table 3.3 gives the skew of the snap-off distribution as +3.10, so a symmetric fence placed on raw values treats the upper tail as anomalous when that tail is simply what a log-normal distribution looks like. On the logarithm the distribution is near-symmetric and the criterion behaves as intended.

For breakthrough pressure the consequence is larger. A mean ± 2 SD rule on raw values would remove the record at 343.5 bar, which is the strongest seal in the dataset and among the most valuable records it contains. A raw-scale outlier removal would delete the rock of most interest to this thesis.

Applied to the logarithm, the rule flags no outlier in the breakthrough population. Its fence runs from 0.14 to 14,100 bar, wider than the measured range of 0.9 to 343.5 bar, because a target this dispersed produces a large interquartile range and a correspondingly permissive fence.

In the snap-off population it flags two, both at 0.2 bar and both the lowest values in the set, one from Egermann et al. (2006) and one from Hildenbrand et al. (2004). Both are retained as measured. Each was traced to its published table during the verification of Section 3.2, so neither is a transcription error, and removing them would discard the low end of a distribution the models are meant to span. They are flagged and not deleted, so their influence can be checked against the delivered dataset.

The precision of the targets is bounded by the pressure step of each step-by-step test, which few studies report. A test raising pressure in 5-bar increments cannot locate breakthrough to better than about 2.5 bar, and this unreported resolution is part of the scatter that no outlier rule can separate from real variation.

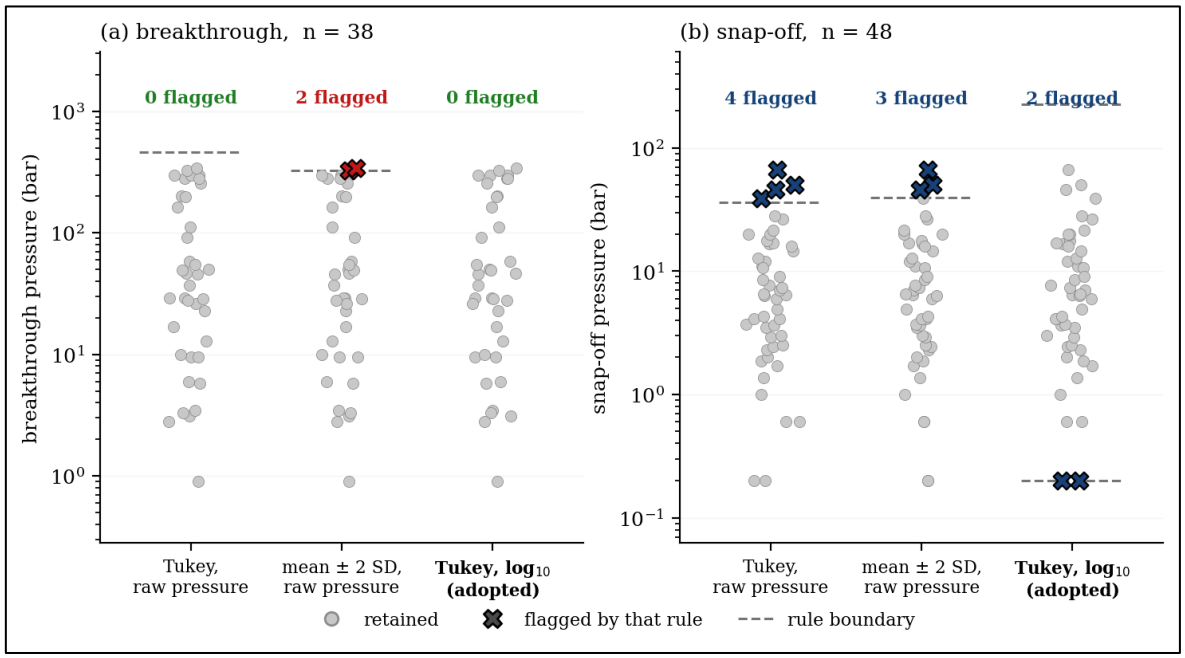


Figure 3.11 — Three outlier criteria applied to the same data.

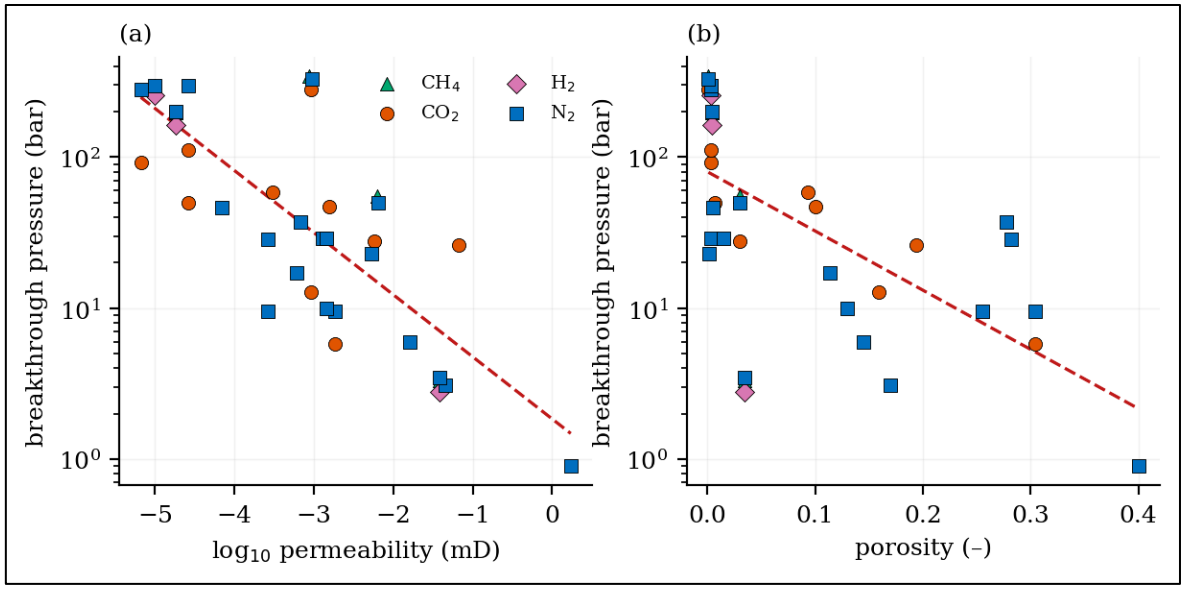


Figure 3.12 — Breakthrough pressure against permeability and porosity.

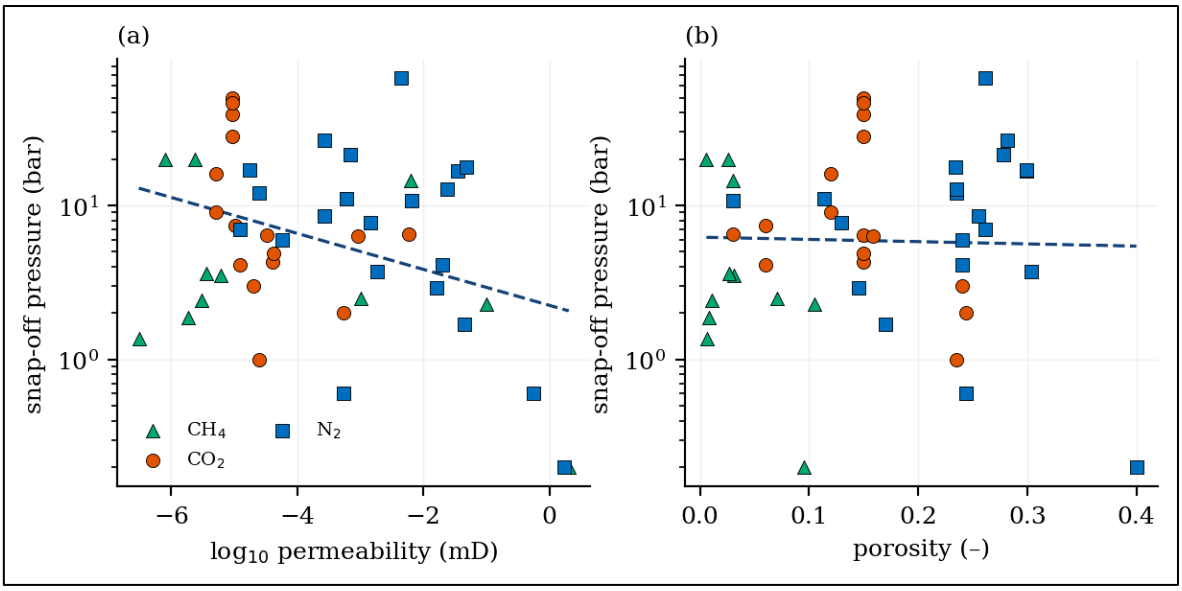


Figure 3.13 — Snap-off pressure against the same two features. The porosity panel is flat.

Figure 3.8 gives the ratio of the median snap-off to the median breakthrough pressure by fluid: 0.26 for carbon dioxide, 0.29 for nitrogen and 0.02 for methane, with a pooled value of 0.19. Figure 3.14 gives the ratio on the thirteen plugs that carry both measurements, and there the median is 0.48.

The two quantities measure different things. A ratio of pooled medians compares one set of rocks measured for breakthrough against a different set measured for snap-off. A same-plug ratio holds the rock fixed and changes only the procedure. The two populations overlap on thirteen plugs only, ten tested with nitrogen (one of them also with carbon dioxide), two with carbon dioxide alone and one with methane, so most of the records behind each by-fluid median contribute to one pressure and not the other. The methane ratio of 0.02 is therefore an artifact of the comparison. Most of its breakthrough values come from outcrop shales and anhydrite, and most of its snap-off values from pelitic mudrocks in a different study; the one methane plug that carries both measurements gives a ratio of 0.26.

The same-plug figure is the one that describes the physics. Thirteen pairs give a median of 0.48, inside the 20 to 50 % range the literature reports, and the individual ratios run from 0.22 to 0.92. The ratio therefore cannot be used as a constant to convert one pressure into the other, and Chapter 6 returns to how snap-off should be bounded instead.

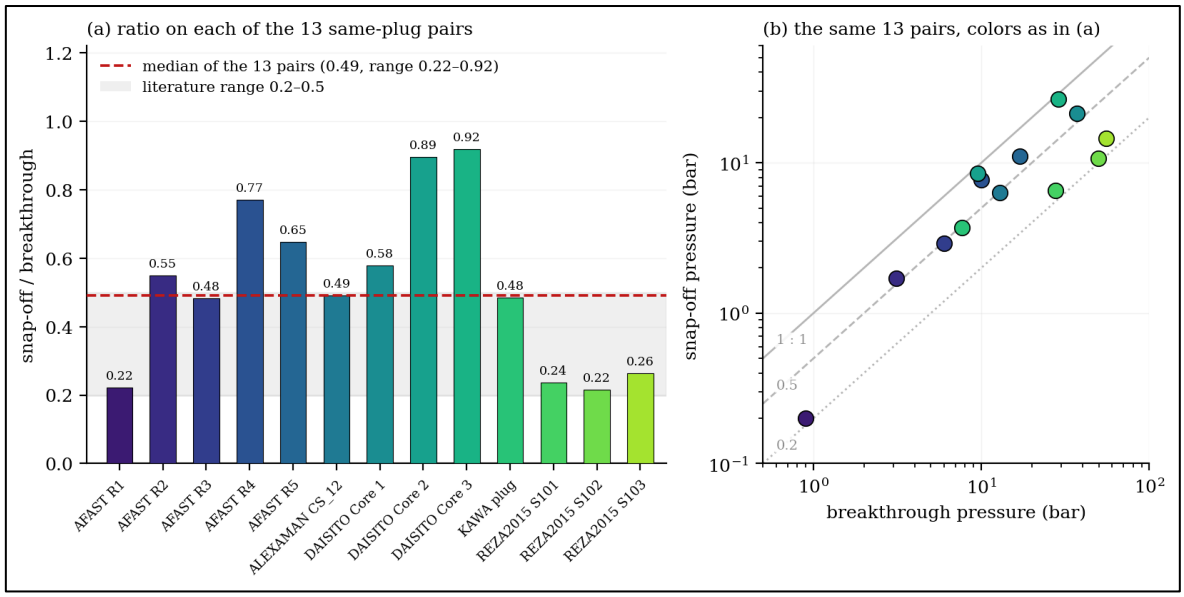


Figure 3.14 — Snap-off relative to breakthrough pressure on the thirteen same-plug pairs.

Decomposed by source, 37 % of the variance of snap-off pressure lies between studies and 63 % within them. Snap-off pressure varies among samples of similar porosity and permeability inside a single laboratory, which places its control in something the compilation does not record. Pore geometry is the explanation the pore-scale literature supports, although within-study measurement scatter cannot be separated from it here.

Each study spans a wide range of its own, and the study medians overlap. This difference is what the pore-scale physics predicts. Breakthrough pressure is set by throat size, which porosity and permeability jointly constrain, so both inform it. Snap-off pressure is set by the pore-body-to-throat aspect ratio, a geometric property that bulk measurements do not resolve, so porosity carries no signal for it. The data therefore validate the physical basis on which the two pressures were separated.

3.9 Within-study and between-study variance

Decomposing the variance of each quantity into a between-study and a within-study part measures how far a quantity is governed by which study reported it.

Table 3.7 — Between-study share of variance.

quantity	***between-study share***
$\log_{10}$ breakthrough pressure	*41.1 %*
$\log_{10}$ snap-off pressure	*36.6 %*
$\log_{10}$ permeability	*33.5 %*
porosity	*85.9 %*

Porosity is by far the most source-structured feature: eighty-six per cent of its variance lies between studies instead of within them. This is a direct consequence of how the compilation is built, since

each study works on rocks from one formation. It bounds how transferable any model trained on compiled data can be, and it is the reason the validation of Chapter 5 holds whole studies out instead of splitting records at random.

3.10 Feature selection

The capillary relation points first to pore-throat size, with specific surface area closely related. Neither is reported in the dataset, so permeability and porosity, the two properties that stand in for pore geometry, are retained as the petrophysical features. They are kept as separate inputs and not collapsed into a Winland-type or flow-zone index, since such indices cannot distinguish a porous low-permeability clay from a tight low-porosity evaporite.

Confining pressure is not retained, although the physics argues for it. Higher effective stress narrows the conducting throats and raises entry pressure, so the property is physically relevant. Three considerations override that.

First, in this dataset confining pressure is a study-level constant: most publications apply a single value to all their samples, so the variable is partly a proxy for source identity and not a stress effect.

Second, requiring it is expensive, and expensive in a way that matters. A 50 to 220 bar window would keep 27 of the 38 breakthrough records and 7 of the 8 sources, counting the records tested without an applied stress as inside it; if an independent confining pressure is also required, 18 records from 5 sources remain. The records it discards are at the high-pressure end of the population. The eleven records outside the window carry porosities from 0.001 to 0.030 and breakthrough pressures from 27.6 to 343.5 bar, and they include every record above 297 bar. Applying the window would truncate the top of the target range from 343.5 bar to 297 bar.

The records the window removes include the tightest rocks and strongest seals in the dataset, which are the rocks a caprock assessment is most concerned with. Excluding them to control a variable whose contribution the data cannot separate from study identity would remove the most relevant part of the population for a gain that cannot be interpreted.

Third, what it adds cannot be attributed. Chapter 5 shows that adding the variable raises the leave-one-source-out coefficient from −0.02 to +0.36 on a fixed population of 24 records. That gain is difficult to read as physics, because in this compilation the variable behaves almost as a source label: of the six studies that report an independent confining pressure, five apply a single value to every sample, and six of the seven distinct values occur in one study only. A model given the confining pressure of a record is close to being given the name of the laboratory that produced it. The confining pressures of a further study are not counted here, because they were set at the injection pressure plus 5 MPa (Section 3.3) and so carry the target itself.

Source-aware validation does not remove this confound. It prevents a model from recognising a study it has already seen; it does not prevent it from learning an association that holds between studies, such as laboratories working at high stress testing tighter rocks. Such an association transfers to a withheld study without being physical.

The data cost reinforces the decision without being its basis. Requiring an independent confining pressure removes fourteen records and two sources, and requiring the stress window that makes those records comparable removes six more, including every record above 297 bar.

The retained feature set is therefore $\log_{10}$ permeability and porosity. Confining pressure is retained in the dataset, and its contribution is examined explicitly as an ablation in Chapter 5 and not assumed.

Two studies sit outside the window entirely, one at ambient conditions and one between 241 and 379 bar.

Depth is not retained because it carries no usable relationship to the laboratory conditions. Temperature is not retained on two grounds. It acts on sealing through interfacial tension, which the dataset carries only as a per-study constant. And the literature does not agree on the sign of its effect: two recent studies on comparable rocks report breakthrough pressure falling and rising with temperature respectively. Interfacial tension and contact angle are not retained as inputs because they are study-level defaults, and lithology is not retained because the label does not fix the rock.

3.11 Petrophysical grouping of the samples

The lithology label is a weak descriptor: permeability within a single label spans several orders of magnitude. A grouping drawn from the measured properties is needed to describe the dataset in sealing-relevant terms. K-means clustering on porosity and log-permeability provides it, applied to the 116 verified records reporting both properties.

Choosing the number of groups. The silhouette score alone does not settle the choice. It rises from 0.462 at two clusters to a maximum of 0.551 at six and falls to 0.499 at seven, so a criterion based on separation alone would favour six groups. The binding constraint here is sample size, and two diagnostics express it.

Table 3.8 — Cluster diagnostics against the number of groups.

k	silhouette	smallest cluster	share of data	fewest breakthrough records in any cluster
2	0.462	45	38.8 %	10
3	0.497	37	31.9 %	6
4	0.517	14	12.1 %	4
5	0.531	14	12.1 %	0
6	0.551	2	1.7 %	0

The smallest cluster holds 37 records at three groups and falls to 14 at four, a fall of nearly two thirds for a silhouette gain of 0.020. The target coverage falls with it. At three groups every cluster carries at least six breakthrough measurements, which is enough to state a median with its spread. At four the sparsest cluster carries four, and at five it carries none. A cluster with no breakthrough record cannot be given a sealing character at all, which is the purpose of the grouping.

Three groups are therefore adopted, at 41, 37 and 38 samples. The choice is set by what 116 records and 38 breakthrough measurements can populate, and not by the separation score, which would favour more.

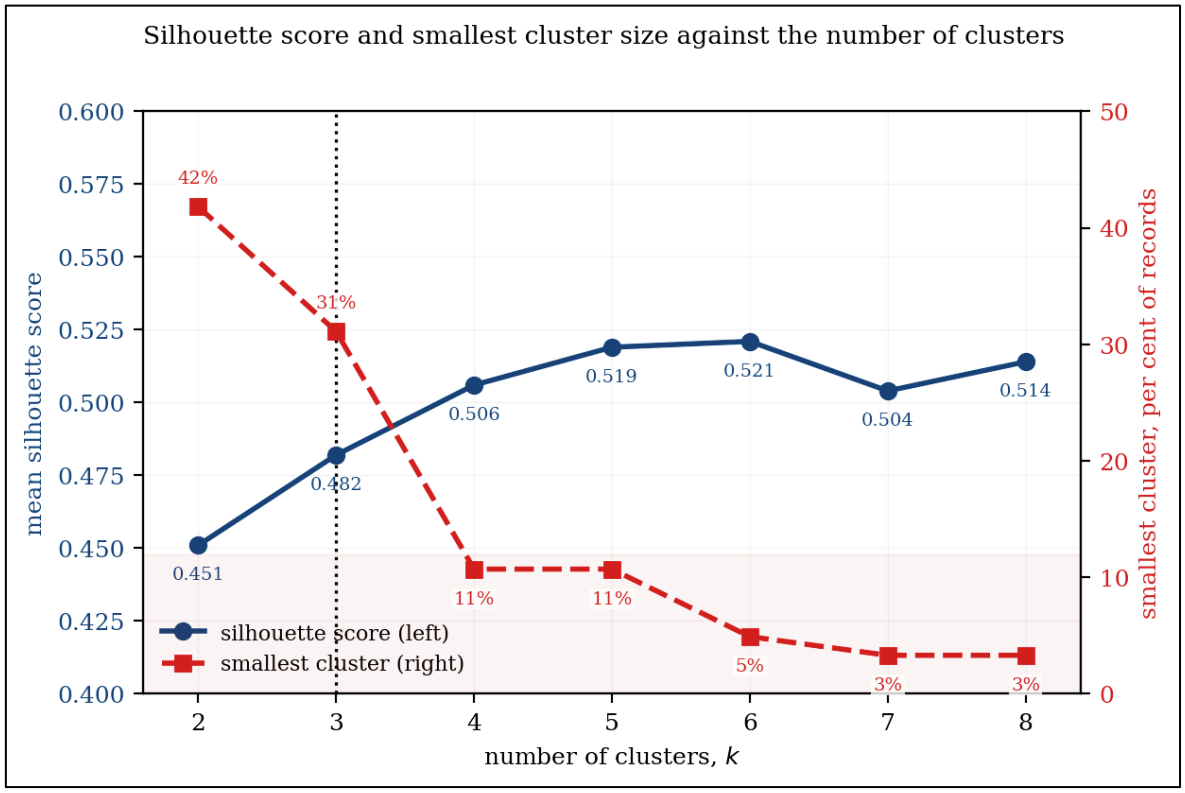


Figure 3.15 — Silhouette score and smallest cluster size against the number of clusters.

The same diagnostics show how the lithology labels perform as a partition of the same space. The silhouette of the 13-label lithology partition in the porosity–permeability plane, with label case normalised and basin annotations removed, is −0.04, against +0.497 for the three k-means groups. A negative value means that a sample is, on average, closer to rocks carrying other labels than to rocks carrying its own. The Adjusted Rand Index between the two partitions is 0.24, so they agree only modestly better than chance. The index is depressed by the mismatch between 13 labels and three groups, which is why the negative silhouette is the more robust of the two statements.

The grouping does not appear to depend on the mercury block. Re-clustering the flow-through and residual records alone, 85 records, gives a silhouette of 0.471 at three groups with a smallest cluster of 26 records. The three-group structure survives removal of the mercury-injection samples, and the same size constraint appears: at five groups the smallest cluster falls to eleven records.

Table 3.9 — The three seal-quality groups.

statistic	*ultra-tight*	*clay-rich*	*moderate*
records	*41*	*37*	*38*
median permeability (mD)	1.8×10^{-5}	6.8×10^{-4}	3.1×10^{-3}
median porosity	*0.021*	*0.278*	*0.035*

statistic	*ultra-tight*	*clay-rich*	*moderate*
median breakthrough (bar), n	*200.0, 12*	*9.5, 6*	*27.0, 20*
median snap-off (bar)	*6.9*	*7.0*	*6.3*

The three groups separate along both features, and the breakthrough medians are informative about which feature does the work. The ultra-tight group, at a median permeability of 1.8×10^{-5} mD and porosity 0.021, has the highest breakthrough pressure at 200 bar. The clay-rich group has an intermediate permeability but by far the highest porosity, at 0.278, and the lowest breakthrough pressure of the three, 9.5 bar, below the moderate group, whose permeability is higher. Breakthrough pressure is therefore not monotonic in permeability across these groups, and the group that breaks the ordering is the one distinguished by porosity. This is the porosity control of Section 3.7 visible at the level of rock classes. The clay-rich group here is high in porosity but low in breakthrough pressure, the opposite arrangement to the clayey facies of Kim et al. (2025), which seal better than the sandy facies of the same formation at equal or slightly lower porosity, a contrast that reinforces, rather than reverses, the negative porosity sign found here.

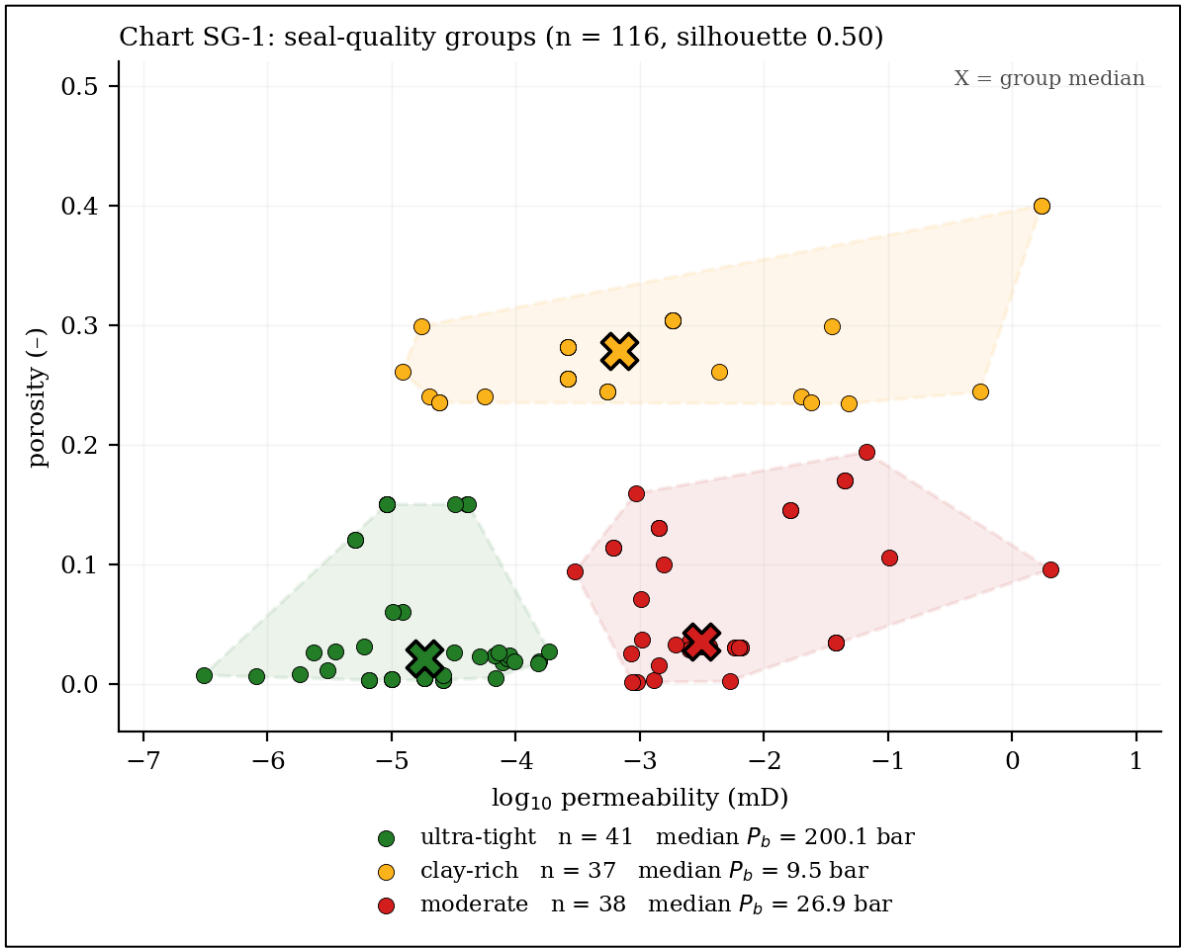


Figure 3.16 — Chart SG-1: the three seal-quality groups in the porosity–permeability plane.

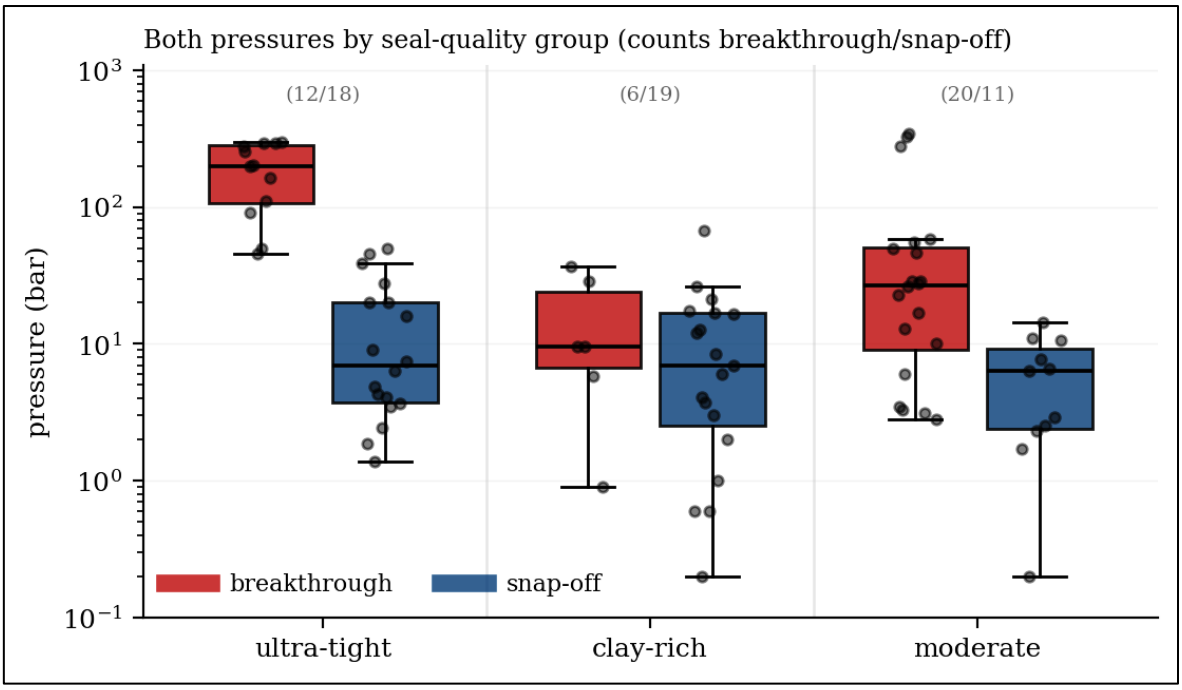


Figure 3.17 — Breakthrough and snap-off pressure by seal group.

The groups are not lithological classes. The ultra-tight group is led by limestone and anhydrite, the clay-rich group by mudstone, clay and mudrock, and the moderate group by shale and carbonate. Each holds several labels, and the same label appears in more than one group.

3.12 Dataset limitations

The analysis rests on a dataset whose limits are quantified here and not left implicit.

Sample size and source concentration. The complete-case requirement leaves 38 breakthrough and 48 snap-off records from 8 and 9 sources. With two features and eight sources, the claim the dataset can support is comparative and not one of absolute accuracy.

Provenance. Nearly two thirds of the records in the merged compilation did not survive verification. That is a result in itself, and it is reported as one. The exclusions are listed with their grounds in Table 3.2 and every retained record carries a DOI and a verification status.

Missing properties. No pore-geometric descriptor is available, which is the most likely reason snap-off pressure resists prediction. The fluid-pair properties are literature constants and not measurements. No quantitative mineralogy is present. Water saturation, which the literature shows raises breakthrough pressure exponentially, is not recorded by any source. Sample length, which one study shows scales linearly with breakthrough pressure, is likewise unrecorded and therefore uncontrolled across the compilation.

Hydrogen. Three hydrogen records enter the breakthrough population, from one study. A second hydrogen study is present in the dataset but excluded from modelling because it does not report the petrophysics. The hydrogen result of Chapter 6 therefore rests on a small and single-sourced sample and is reported as indicative instead of established.

Repeat measurements. Twenty-six plugs carry more than one record, from repeat runs, from tests with different fluids, or from a step-by-step and a residual test on the same plug. Where a study reports successive measurements on one plug the values are not independent, and the Plug_Id attribute allows this to be checked. One source show sealing capacity falling by a third to a half over successive exposures to carbon dioxide, so repeat measurements on one plug are not replicates of a fixed quantity. Two plugs in the largest source carry a high-pressure-side leakage note from the original study, and the record whose value enters the modelling is flagged in the dataset; the flag is documentary, since leakage after breakthrough affects the effective permeability and not the residual pressure the source reports.

All statistics and groupings in this chapter were computed with Python 3.12, scikit-learn, SciPy and pandas, with k-means run at a fixed random seed of 42 and ten initializations, so that every figure and table can be regenerated from the delivered dataset.

Within these limits the dataset serves its purpose. Part I consolidated the compilation, verified every record against its source, added five studies, and reduced 413 records to 146 verified core-scale measurements. Part II established that the two sealing pressures behave as the physics requires, that porosity and permeability both inform breakthrough pressure while porosity carries no signal for snap-off pressure and permeability only a weak one, and that a data-driven grouping describes the rocks better than the lithology labels.

Chapter 4 — Charts for Estimating Breakthrough Pressure

A full model estimate of breakthrough pressure needs the two properties Chapter 3 retained, permeability and porosity. A field assessment rarely has both to hand, and often has only a permeability, a porosity, or a mercury-injection test. This chapter presents two charts that give a first estimate from those measurements alone, in the manner of a log-interpretation chart. The first reads breakthrough pressure from permeability and porosity. The second converts that value between storage fluids, and converts a mercury-injection entry pressure into a breakthrough estimate. The two are used in sequence, and each states the uncertainty a reader should attach to the answer.

4.1 Chart BP-1: breakthrough pressure from permeability and porosity

Breakthrough pressure follows the Thomas (1968) form, a decreasing power of permeability, with porosity as a second control. Fitting that form to the 38 flow-through breakthrough records of the verified compilation gives

$$\log_{10} P_b = 0.696 - 0.331 \log_{10} k - 2.082\,\varphi \qquad (4.1)$$

with P_b in bar, k in millidarcy and φ as a fraction, and a coefficient of determination of 0.70.

Chart BP-1 (Figure 4.1) draws this relation as a fan of constant-porosity curves. Permeability is entered on the bottom scale, the curve for the sample porosity is followed upward, and the breakthrough pressure is read on the left scale. The shaded band is one standard deviation of the data about the fit, a factor of about 2.4, and it is the uncertainty to attach to a reading.

Two points affect how the chart is read. The relation is fitted to the same 38 records the models of Chapter 5 use, so the chart and the model are the same relationship written two ways. What separates them is not the relationship but what is known about how far it transfers. Section 6.6 takes that up. And the porosity term matters: over the porosity range the compilation spans, from 0.001 to 0.400, it moves the estimate by a factor of nearly seven, which is why a permeability-only reading is insufficient.

Table 4.1 — Breakthrough pressure read from Chart BP-1, in bar, at four porosities.

permeability (mD)	φ = 0.02	φ = 0.05	φ = 0.10	φ = 0.30
10^{-6}	436	378	298	115
10^{-5}	204	176	139	54
10^{-4}	95	82	65	25
10^{-3}	44	38	30	12
10^{-2}	21	18	14	5.4
10^{-1}	9.7	8.4	6.6	2.5

Each column is Equation 4.1 evaluated at that porosity, and the band on any reading is a factor of 2.4. Two features are worth reading off the table. Across four orders of magnitude in permeability the estimate falls by a factor of about 21, and across the porosity range from 0.02 to 0.30 it falls by a factor of about 3.8 at fixed permeability. The porosity term is therefore substantial, and a permeability-only reading can be wrong by more than the uncertainty band.

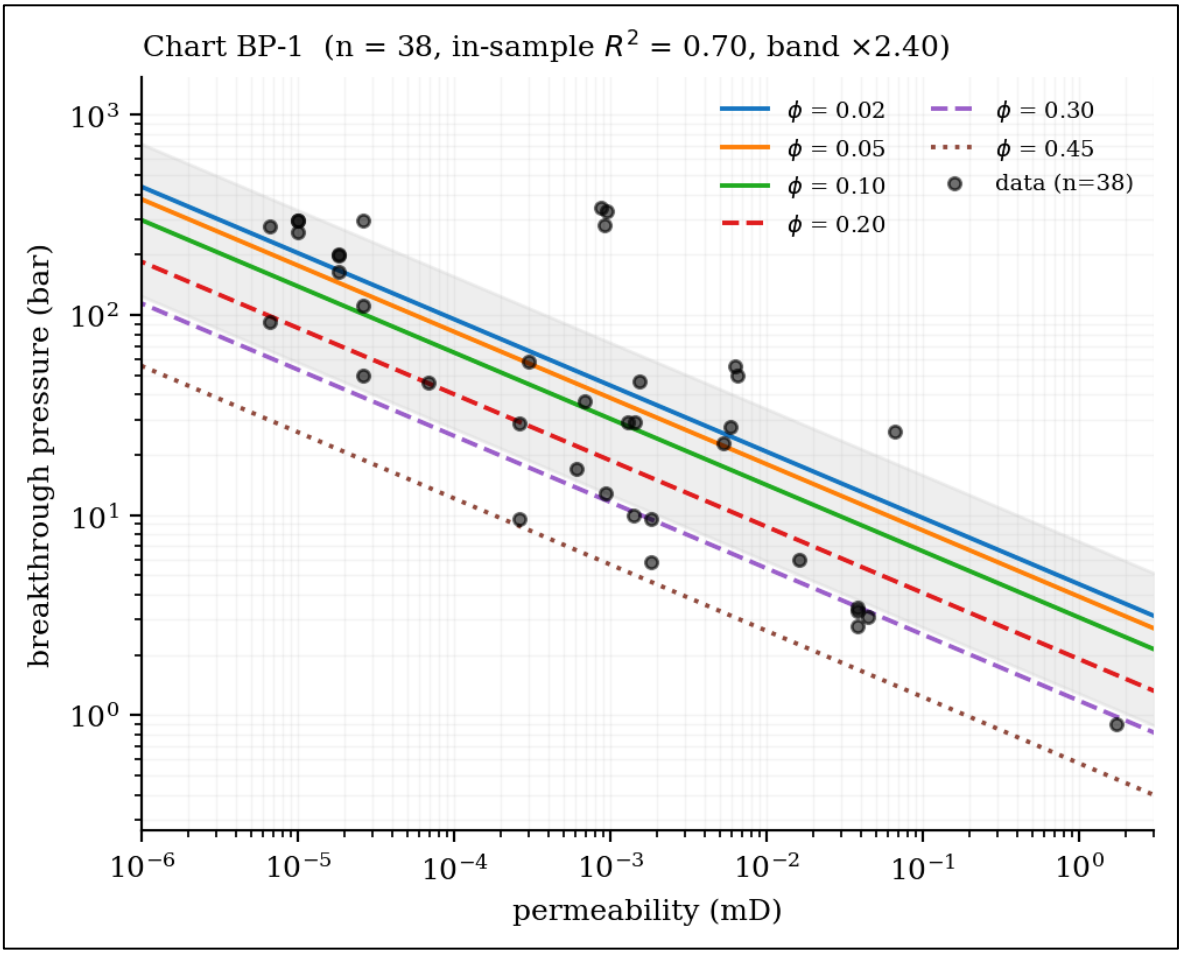


Figure 4.1 — Chart BP-1: breakthrough pressure from permeability and porosity.

The chart gives the breakthrough pressure of the fluid the sample was tested with, which for most of the compilation is carbon dioxide or nitrogen. Moving that value to another storage fluid, or reading it from a mercury test, is the task of the second chart.

4.2 Chart BP-2: fluid conversion and mercury injection

A mercury-injection test measures the pressure at which mercury enters the pore throats of a dry sample against air. A breakthrough test measures the pressure at which a gas enters the same throats against brine. Both are the capillary entry pressure of the same pore geometry, so one follows from the other through the ratio of the two capillary systems:

$$P_b = CF \cdot P_{Hg}, \qquad CF = \frac{\sigma_{gas}\cos\theta_{gas}}{\sigma_{Hg}\cos\theta_{Hg}} \qquad (4.2)$$

With mercury–air values of 485 mN/m and 140°, and the storage-fluid values carried by the source studies at a zero contact angle, the factors are:

Table 4.2 — Mercury-to-gas conversion factors (Equation 4.2, ϑ_gas = 0°).

fluid	σ (mN/m)	CF
N_2	72	0.194
CH_4	72	0.194
H_2	71.6 (51–72)	0.193 (0.138–0.193)
CO_2	60	0.161

The same ratio converts a breakthrough pressure from one fluid to another: divide by one fluid factor and multiply by the other.

The carbon-dioxide factor uses 60 mN/m, the low-pressure value carried by most of the source studies. At storage conditions carbon dioxide is supercritical and its interfacial tension against brine is much lower (Section 2.5). The studies in the compilation that worked at high pressure carry 27 mN/m and 34°, which give a factor of about 0.06. For carbon dioxide at storage conditions that lower factor is the one to use, and it reduces the estimate by a factor of between two and three.

The hydrogen line is less well constrained than the other three, and the range in Table 4.2 records that. Li, Gates & Hejazi (2025) measured the hydrogen–brine interfacial tension directly at 71.6 mN/m, within a per cent of the nitrogen and methane values, and the line is drawn at that value. Al-Mukainah et al. (2022), measuring on brine at 50 °C, report 63.7 mN/m at atmospheric pressure falling to 51.3 mN/m at 69 bar, which by Equation 4.2 gives a factor of 0.14 to 0.17 instead of 0.19. The two measurements do not agree, and the compilation contains no hydrogen record that decides between them. On the one study that measured breakthrough pressure for hydrogen, nitrogen and methane on the same plugs, hydrogen broke through 14 to 19 per cent below nitrogen (Section 3.3), an offset of the size the lower interfacial-tension values would predict and also within the scatter of a single step-by-step test. A further study reports hydrogen contacts angles on mica consistently lower than carbon dioxide, which by Equation 4.2 implies greater capillary resistance to hydrogen, not less. The hydrogen line is therefore drawn at the Li value, and a reading from it should be taken as an upper estimate, with the lower factor of 0.14 as the other end of its range.

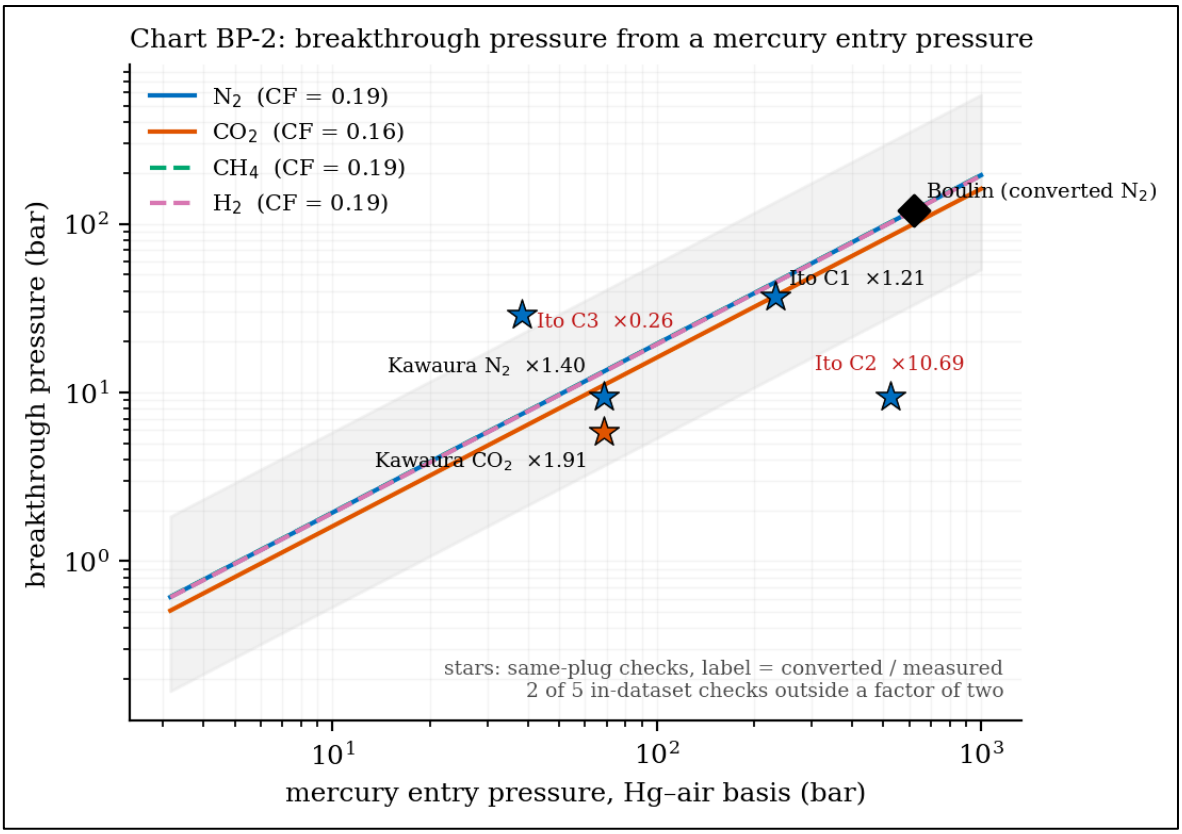


Figure 4.2 — Chart BP-2: breakthrough pressure for each storage fluid from a mercury-injection entry pressure. The mercury value is read on the bottom scale and the fluid line followed to the left scale; symbols are the same-plug checks.

The contact angle. The zero value is the conventional assumption for the mercury conversion. It is not the value the same studies use for their gas tests, and the choice matters: at a contact angle of 60° each factor roughly halves. The zero value is kept so that the estimates are comparable with the published literature. Independent support for the 60° figure exists: one study back-calculates contact angles from the breakthrough measurements of two of the sources used here and recovers cos θ of about 0.4 to 0.6, which it describes as upper bounds, that is θ near 60°, the value the compilation carries for those records; the source studies themselves assumed 0° in their pore-size calculations.

The consequence is large. Section 2.5 describes a basin-scale study in which the maximum retainable CO_2 column falls from more than 1000 m, under a fixed water-wet angle, to zero below about 2200 m once the angle is allowed to shift with depth. No measurement in the present compilation resolves which assumption to prefer.

The saturation convention. A mercury entry pressure is read from a curve, and which point on that curve defines the entry pressure is a convention, not a measurement. Studies in this compilation use at least three: the pressure at ten per cent mercury saturation, the pressure at twenty per cent, and the first detectable intrusion. These give materially different numbers from the same sample. Chapter 3 documents one source where the first-intrusion reading is one to two orders of magnitude below the twenty-per-cent reading, because the first mercury to enter a plug fills surface irregularities and not the pore network. This is a second source of uncertainty in the mercury

conversion, additional to contact angle and of comparable size. The mercury records used here are placed on the twenty-per-cent convention for that reason.

4.3 Validation of the mercury conversion

Five plug–fluid comparisons in the compilation carry both a mercury value and a flow-through breakthrough measurement on the same plug, which is the only way to check the conversion without the rock changing between the two tests. Table 4.3 lists them.

Table 4.3 — Same-plug checks of the mercury conversion in the compilation (Equation 4.2, ϑ_gas = 0°).

plug	fluid	mercury–air entry pressure (bar)	converted (bar)	measured step-by-step (bar)	converted ÷ measured
Ito et al. (2011), Core 1	N_2	231.3	44.9	37.0	× 1.21
Ito et al. (2011), Core 2	N_2	523.7	101.6	9.5	× 10.7
Ito et al. (2011), Core 3	N_2	38.4	7.4	28.75	× 0.26
Kawaura et al. (2013)	N_2	68.7	13.3	9.5	× 1.40
Kawaura et al. (2013)	CO_2	68.7	11.1	5.8	× 1.91

The two Kawaura points are checks on fragments of a single mudrock core rather than on separate plugs, and Kawaura et al. (2013) report the step-by-step nitrogen threshold as a range (0.9 to 1.0 MPa), so its entry is read at the range midpoint. Three of the five fall within a factor of two. Two fail, and they fail in opposite directions. Core 2, a mudstone, converts to about 100 bar against a measured 9.5 bar, an over-estimate by a factor of ten. Core 3 converts to 7.4 bar against a measured 28.75 bar, an under-estimate by a factor of almost four. Their source study attributes the disagreement to heterogeneity: cuttings from the same seven-centimetre depth interval gave mercury values spanning a factor of seven, so the pore structure is strongly heterogeneous and a single mercury reading does not characterise it.

The conversion can fail in both directions, which is the firmest conclusion these data support. One study found a measured supercritical CO_2 value twice the value recalculated from its own nitrogen measurement through Equation 4.2, and suggested either that published CO_2–brine interfacial tension is underestimated or that interfacial tension and contact angle have a smaller influence on threshold pressure than the Laplace relation implies. Another, working on a carbonate-rich caprock, found nitrogen requiring 4.5 to 6 MPa while CO_2 broke through the same plug below 1.5 MPa. That is a factor of more than three where the interfacial tensions differ by less than two, with no wettability alteration detected, and the authors call the result contrary to expectation.

The wider literature reaches the same conclusion for fine-grained rocks. One study measuring both mercury-injection and gas displacement pressures on a suite of mudstones and clays found no

correlation between the two, attributing the mismatch to the anisotropy of mudrocks and to the drying the mercury method requires. Another judged mercury injection unsuitable for an accurate breakthrough pressure, noting that mercury injection is isobaric at zero effective stress while the entry pressure depends on the stress state.

The mercury conversion is therefore an order-of-magnitude estimate. It is useful for a first assessment of a caprock from a test that takes hours instead of weeks, and it is not a substitute for a flow-through measurement where sealing capacity must be established.

4.4 Application of the charts

The two charts answer the questions a field assessment is likely to ask. Where permeability and porosity are known, Chart BP-1 gives the breakthrough pressure directly, with the factor-2.4 band as its uncertainty, and Chart SG-1 (Figure 3.16) places the sample in a seal-quality group. Where only a mercury-injection test is available, Chart BP-2 gives the estimate from the mercury value, with a band of at least a factor of three, which two of the five same-plug checks exceed. Where a breakthrough pressure is known for one fluid and wanted for another, Chart BP-2 converts between them.

A worked example is traced on both charts. A sample of permeability 5×10^{-4} mD and porosity 0.10 reads a breakthrough pressure of about 38 bar on Chart BP-1. A mercury entry pressure of 150 bar on the same sample, read at the twenty-per-cent convention, gives about 29 bar for nitrogen on Chart BP-2. The two estimates agree within the bands.

Four limitations bound both charts. They are fitted to 38 breakthrough records and checked against five same-plug mercury comparisons, two of which fail, so they are guides to the order of magnitude and not exact predictions. The mercury conversion assumes a zero contact angle, and a more water-repellent rock lowers the estimate by up to a factor of two. The saturation convention behind any mercury reading must be known, since the conventions in use differ by more than the fluid conversion does. And the hydrogen line rests on three breakthrough records from one study and on two interfacial-tension measurements that disagree by up to 30 per cent, so a hydrogen reading carries a further factor of up to 1.4 of its own.

Within those limits the charts turn the measurements most readily available for a caprock into a breakthrough estimate for any of four storage fluids, with a stated estimate of how far it can be trusted.

Chapter 5 — Predictive Modelling under Source-Aware Validation

Chapter 3 established what the data can support; breakthrough pressure correlates with permeability and porosity as the capillary physics requires, while snap-off pressure carries no porosity signal and only a weak permeability one. Chapter 4 turned the first of those relationships into a chart that needs no computer. This chapter puts the same relationship to a harder test. Four models are fitted to the same data, from the classical engineering correlation to two machine-learning ensembles. They are compared under a validation scheme built around the heterogenous nature of the geology.

The chapter sets out the prediction task and the reason fluid properties are rejected, the four models compared, the source-aware validation protocol, and the comparison itself.

5.1 The prediction task

The problem is supervised regression on a continuous target. Each record supplies two inputs, the base-ten logarithm of permeability and the porosity, and the model returns an estimate of the sealing pressure.

Fluid properties are absent from that input set because of the data. Sealing pressure depends on the fluid–rock interaction through the interfacial tension and the contact angle, as the Young–Laplace relation makes explicit. In this compilation both quantities appear only as study-level literature constants: a single value applied to every sample of a study, not a measurement on the individual rock–fluid pair. A property that takes one value per study cannot inform a model at the sample level, and entering it would allow the model to read the study label in place of a physical effect. The fluid effect the physics requires is not ignored; it is handled after the rock term, through the interfacial-tension conversion of Chapter 4.

Confining pressure is left out by choice, as argued in Section 3.10 and tested later in this chapter. The variable is physically relevant, but in this dataset, it is close to a study label; requiring an independent value costs fourteen of the thirty-eight breakthrough records and two entire studies; and on a fixed population it raises the score by an amount that cannot be attributed to stress; the numbers follow below.

Two decisions about the target follow from the physics. First, the target is modelled as its base-ten logarithm. Sealing pressures span roughly three orders of magnitude, and the Young–Laplace relation is multiplicative in the pore-throat radius, so the error of a prediction is naturally proportional to its value. Working in $\log_{10}$P makes a prediction accurate in the sense of a ratio to the measured value,

$$\varepsilon = \log_{10}\left(\hat{P}/P\right) \qquad (5.1)$$

so that an error of 0.30 in log units is a factor of two in pressure, whichever part of the range it falls in. Every coefficient of determination in this chapter is computed on $\log_{10}$P. Second, breakthrough and snap-off are fitted as two separate problems with separate training sets, because Chapter 3 showed them to be governed by different predictors. No model is fitted to the two pooled together.

Permeability enters as its logarithm for the same reason as the target: it spans about seven orders of magnitude, and on a linear scale the ultra-tight seals, the samples of most interest, would collapse onto a value indistinguishable from zero. Porosity enters untransformed.

The pressure range needs a comment before any result is read. On the working convention of Section 1.5, a caprock with a breakthrough pressure below about 20 to 30 bar is a poor seal, and one below 10 bar would not be classed as a caprock at all, the latter corresponding to only a short retainable CO_2 column. The compilation contains records below these values and they are retained deliberately, so that the models are trained and judged across the widest range of sealing behaviour the data offer. Those low-value records are not presented as caprocks; they widen the scenario against which the models are tested.

5.2 Models compared

Four models are fitted to each target. They span the range from the correlation an engineer already uses to flexible ensembles, so that the comparison separates how much of any gain comes from adding inputs and how much from the flexibility of the method.

The Thomas baseline is the permeability-only power law of Chapter 2, refitted to the present data. In logarithmic form the power law is a straight line,

$$\log_{10} P_b = a + b \log_{10} k \qquad (5.2)$$

fitted by ordinary least squares with the intercept and slope refitted instead of fixed at the 1968 constants. The intercept and slope are refitted so that the baseline is the best permeability-only law this dataset supports, not one study's calibration. Fitted here it gives

$$\log_{10} P_b = 0.267 - 0.411 \log_{10} k$$

Refitting matters: the correlation with its 1968 published constants, evaluated unchanged on this population, reaches a coefficient of only 0.10 and a median error of a factor of 3.2, because it was calibrated on a different suite of rocks. The comparison in this chapter is therefore against the refitted law, which is the best permeability-only fit these data support and a far stronger baseline than the published constants.

The multivariate linear regression adds porosity, still as a straight-line relationship,

$$\log_{10} P_b = a + b \log_{10} k + c\,\varphi \qquad (5.3)$$

and fitted here gives

$$\log_{10} P_b = 0.696 - 0.331 \log_{10} k - 2.082\,\varphi$$

It is the simplest model that uses the full input set, and its role is diagnostic: if a flexible method does no better than this, the underlying relationship is close to log-linear and the extra machinery is unnecessary.

The random forest is an ensemble of regression trees, each grown on a bootstrap resample of the rows and each choosing its splits from a random subset of the inputs, with the predictions averaged. Because each tree sees a different resample the individual trees are decorrelated, and averaging reduces the variance that makes any single deep tree unstable. The forest can represent thresholds and interactions without those being specified in advance.

The gradient-boosted regressor also builds trees, but in sequence: the first makes a rough prediction, the second is fitted to the error the first left behind, and so on, each added at a small weight. Boosting can extract more signal than a forest from a small dataset, but for the same reason it will fit noise if allowed to run too deep or too long.

Table 5.1 — Model configurations.

model	*inputs*	*configuration*
Thomas baseline	$\log_{10} k$	*ordinary least squares, refitted intercept and slope*
Multivariate linear	$\log_{10} k$, φ	*ordinary least squares*
Random forest	$\log_{10} k$, φ	*300 trees, maximum depth 4, minimum 3 samples per leaf*
Gradient boosting	$\log_{10} k$, φ	*200 rounds, maximum depth 2, learning rate 0.05, subsample 0.8*

Hyperparameters were fixed a priori from sample size and physical reasoning rather than optimized. With 38 records and two features the data do not support a separate tuning partition, and searching hyperparameters against the same validation estimate later reported would bias it upward. The ensembles are therefore reasonable, not an optimal configuration, a limitation carried forward. No feature scaling is applied to the four core models: the tree ensembles are unaffected by rescaling, and for the linear models scaling changes the coefficients but neither the fit nor the predictions. The support-vector regression and the neural network of the complexity stress-test below, and the k-means grouping of Chapter 3, are the exceptions and are fitted on standardized features, as those methods require.

5.3 The source-aware validation protocol

How a model is validated decides what its reported accuracy means, and for a compilation drawn from many laboratories the choice is not a formality.

Leave-one-source-out validation is the primary estimate. The records are grouped by originating study, and each study is withheld in turn while the model trains on all the others and predicts the withheld one. Every prediction is therefore made on rocks from a study the model has never seen, which is the situation of applying a published model to a new site. The grouping key is the verified source attribution of Chapter 3, which is why that verification mattered: a wrong source label would place a study's records on both sides of the split and inflate the result. The verification found exactly that in one block of twenty-five records, which was removed.

Random five-fold cross-validation (a single shuffled split at the fixed seed of 42) is reported alongside it as a diagnostic, not as a competing measure. A random split puts records from the same

study on both sides of the partition, so a model can score well by recognizing a study it has already partly seen. The difference between the two estimates measures how much a model leans on that recognition.

Because both rest on a small sample, the difference between two models can easily be noise. Uncertainty is quantified by bootstrap resampling of the pooled out-of-fold predictions, 2000 replicates, reported as a 95 % percentile interval. Two versions are reported. The record-level bootstrap resamples the 38 records and treats them as independent, which they are not within a study; the source-level bootstrap resamples the eight studies with replacement and so respects the grouping the validation design rests on. A third, plug-level bootstrap resamples the 29 distinct plugs, and sits between the two: it removes the double-counting of repeat measurements on one plug without collapsing the sample to eight groups. The source-level bootstrap is the most conservative of the three, and it is the one to read when asking whether a model transfers at all.

The tree ensembles and the bootstrap depend on a random seed, fixed at 42 throughout. The dependence is small: re-running the two ensembles at ten other seeds moves the random-forest estimate between +0.22 and +0.25 and the gradient-boosting estimate between +0.25 and +0.29, and re-drawing the record-level bootstrap of the multivariate model at ten seeds moves its lower bound between +0.11 and +0.16 and its upper bound by less than 0.01. None of the comparisons below turns on the seed.

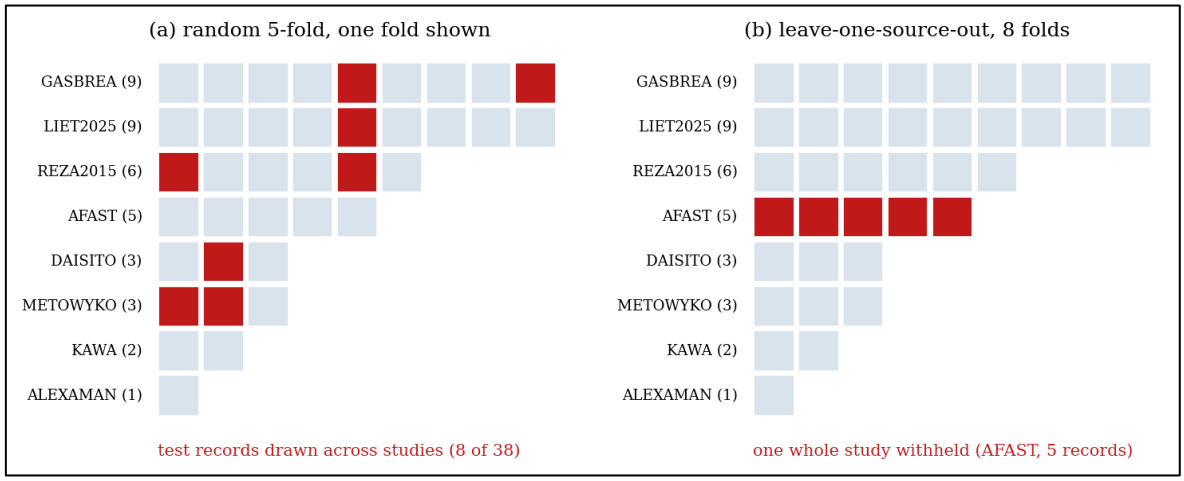


Figure 5.1 — Random and source-aware partitioning, and the fold structure the design produces on the breakthrough population.

Figure 5.1 shows the design and what it means for the present data. The breakthrough population divides into eight folds of very unequal size: the two largest sources contribute nine records each and the smallest a single record, so each fold withholds between 3 and 24 per cent of the data. The estimate is therefore an average over eight unequally weighted tests, which is one reason the bootstrap interval reported below is wide.

A confining-pressure ablation compares the retained formulation against one that includes the variable, separating any accuracy attributable to stress from the part confounded with provenance.

A source-removal test recomputes the estimate with each of the two largest contributing studies withheld entirely.

5.4 Breakthrough pressure results

Table 5.2 and Figure 5.3 give the comparison for breakthrough pressure on the 38-record population from eight studies.

Table 5.2 — Model comparison for breakthrough pressure (n = 38, eight sources, 29 plugs). R^2 computed on $\log_{10}P$; the three intervals are 95 % bootstrap percentile intervals on the leave-one-source-out estimate, resampling records, plugs and whole studies respectively.

model	in-sample fit	random 5-fold	leave-one-source-out	record-level 95 %	plug-level 95 %	source-level 95 %	RMSE (log units)	median error
Thomas baseline	0.62	0.57	+0.46	+0.12 to +0.68	−0.07 to +0.69	−0.71 to +0.78	0.51	× 2.60
Multivariate linear	0.70	0.67	+0.47	+0.15 to +0.68	+0.04 to +0.70	−0.37 to +0.71	0.50	× 1.97
Random forest	0.85	0.74	+0.24	−0.21 to +0.48	−0.34 to +0.50	−0.94 to +0.51	0.61	× 2.64
Gradient boosting	0.97	0.70	+0.28	−0.14 to +0.50	−0.30 to +0.50	−0.94 to +0.59	0.59	× 3.19

The comparison in Table 5.2 supports four findings.

First, every model loses accuracy when it is asked to predict an unseen study. The in-sample fit overstates transferable skill in every case, and the random five-fold estimate overstates it too: the gap between the random split and the source-aware estimate is 0.11 for the Thomas baseline, 0.20 for the multivariate model, 0.50 for the forest and 0.42 for gradient boosting. A validation design that ignores provenance would have reported roughly two fifths again as much accuracy as the linear model can actually deliver, and for the ensembles between two and a half and three times as much.

Second, only the two linear models have a transferable skill distinguishable from zero, and only under the weaker of the two bootstraps. The multivariate model's leave-one-source-out coefficient is +0.47, with a record-level interval of +0.15 to +0.68; the refitted Thomas baseline reaches +0.46 with an interval of +0.12 to +0.68. Both exclude zero, and neither ensemble does. Resampling plugs rather than records removes the double-counting of repeat measurements on one plug. Under this bootstrap the multivariate interval remains just clear of zero (+0.04 to +0.70) while the Thomas interval does not (−0.07 to +0.69), so once within-plug repeats are accounted for the multivariate

model's transferable skill is the more robust of the two. Resampling whole studies widens the multivariate interval to −0.37 to +0.71 and the Thomas interval to −0.71 to +0.78, and on that reading no model in the table can be shown to beat the pooled mean of the population. The multivariate model's RMSE of 0.50 log units corresponds to a factor of 3.2, and its median absolute error of 0.29 log units to a factor of 2.0 of the measured value. The evidence that any model transfers is therefore suggestive, not conclusive, and it is confined to the two simplest models.

The sample of 38 records limits the strength of this claim. The record-level interval on the multivariate model spans 0.53, the source-level interval 1.09, and the intervals of all four models overlap heavily. The Thomas baseline reaches +0.46, within 0.01 of the multivariate model, and its upper bound coincides with it. What the comparison establishes is which models the data support, not that their skill is established beyond doubt. It does not establish that the multivariate model is superior to the Thomas correlation on the coefficient of determination, and Section 6.9 shows that on the median error the two trade places across the pressure range. The ranking among the models is not resolved at this sample size, and no claim is made that it is.

Third, what porosity adds appears in the error scale and not in the coefficient. The multivariate linear model matches the permeability-only baseline in the source-aware estimate, +0.47 against +0.46, and lowers the median error from a factor of 2.60 to 1.97 while leaving the RMSE almost unchanged. Both tree ensembles, which have every opportunity to represent non-linearity and interaction, transfer worse than either straight line. The relationship this dataset supports is close to log-linear in permeability, with porosity as a second term that tightens the typical error without raising the explained variance, and the additional flexibility does not improve transfer.

Fourth, gradient boosting illustrates the problem most clearly. It fits the training data almost perfectly, at 0.97, and transfers at 0.28, a gap of 0.69 among the largest in the table. This is the pattern expected of a flexible learner on a small heterogeneous dataset: it has enough freedom to describe the particular records it was shown, including their noise, and that description does not carry to a new study.

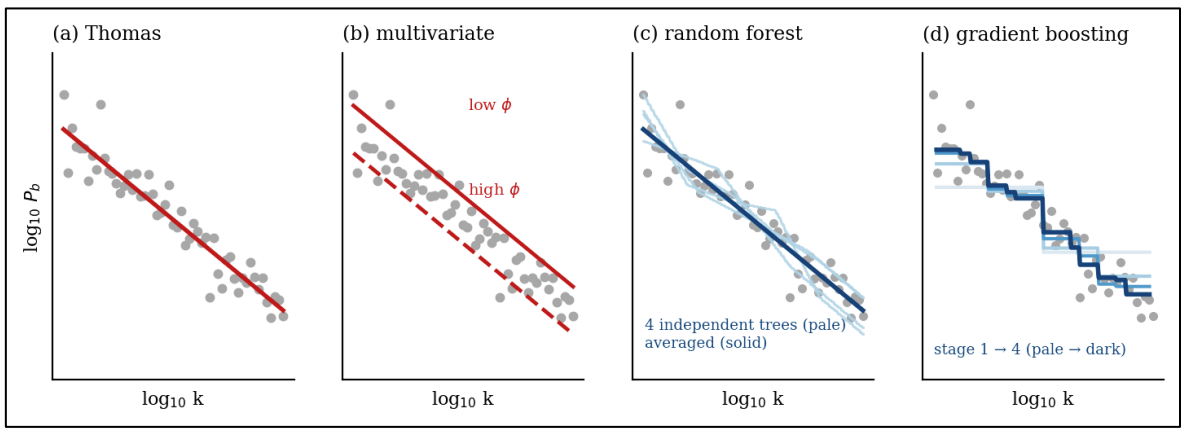


Figure 5.2 — The four model families. The Thomas baseline is one line in log-permeability, the multivariate model adds a porosity term, the forest averages independent trees and boosting corrects its own residual in sequence.

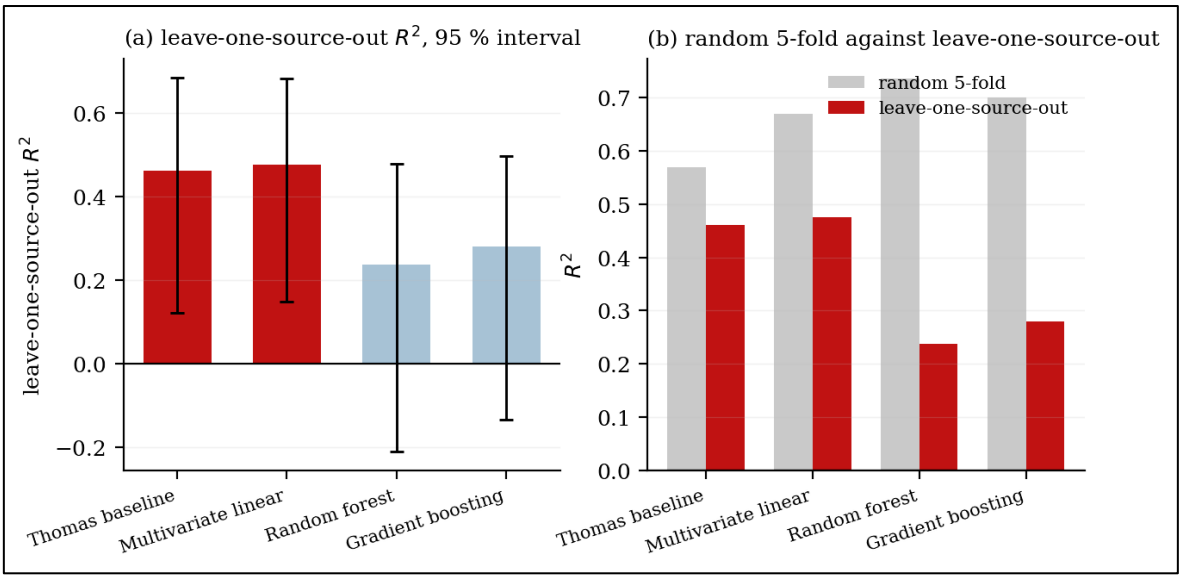


Figure 5.3 — Model comparison for breakthrough pressure: leave-one-source-out R^2 with 95 per cent record-level bootstrap intervals, and the same estimate against the random five-fold estimate.

Figure 5.4 shows the same comparison sample by sample, as predicted against measured pressure, with a band spanning a factor of two either way. The Thomas baseline flattens at both ends, over-predicting the weakest seals and under-predicting the strongest, which is what a permeability-only law does when porosity also matters. The multivariate model follows the line more closely through the middle of the range, though Section 6.9 shows that it shares the same reversal of bias at the two extremes.

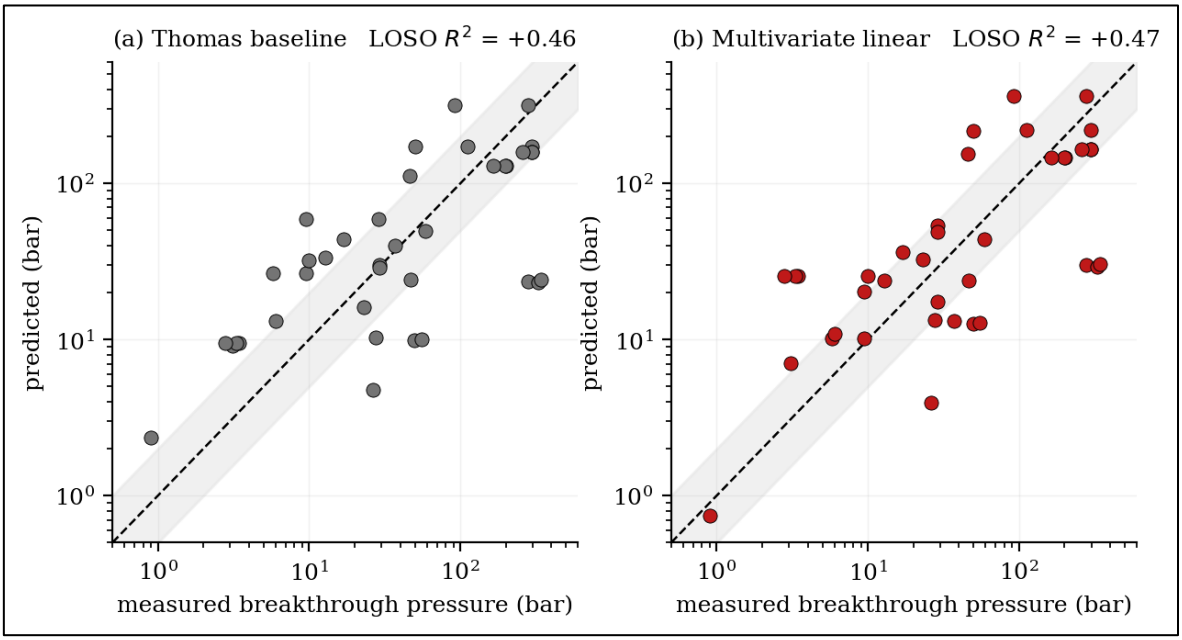


Figure 5.4 — Predicted against measured breakthrough pressure, from leave-one-source-out predictions.

5.5 Effect of model complexity

If the ceiling in Table 5.2 were a limitation of the four methods chosen, a more flexible method would break through it. Three further methods were therefore fitted under the identical protocol: a support-vector regressor with a radial-basis kernel, an AdaBoost regressor and a small feed-forward neural network. They are not proposed as models for use. At this sample size their settings cannot be honestly tuned, and a neural network in particular is inappropriate for tens of records. They test whether flexibility is what the problem lacks, and the results suggest it is not.

Table 5.3 — Complexity stress-test, all seven methods under leave-one-source-out validation.

method		***in-sample***	***LOSO***	***gap***
Multivariate linear	*core*	*+0.70*	*+0.47*	*+0.23*
Thomas baseline	*core*	*+0.62*	*+0.46*	*+0.16*
Neural network	*added*	*+0.79*	*+0.40*	*+0.39*
Gradient boosting	*core*	*+0.97*	*+0.28*	*+0.69*
Random forest	*core*	*+0.85*	*+0.24*	*+0.62*
Support-vector regression	*added*	*+0.83*	*+0.15*	*+0.69*
AdaBoost	*added*	*+0.95*	*+0.10*	*+0.85*

Seven methods spanning linear regression, bagged and boosted trees, kernel regression and a neural network all reach the same modest ceiling. The two best transfer scores belong to the two linear models, which also have by far the smallest gaps between in-sample fit and transfer. AdaBoost fits its training data at 0.95 and transfers at 0.10, the largest gap in the table.

The pattern is consistent: the methods that describe their training data best transfer worst. The limit appears to lie in the data, as Chapter 3 anticipated: thirty-eight records from eight sources, no fluid–rock properties, and a target that varies by orders of magnitude between studies.

Two fixed baselines outside the fitted comparison make the same point from the other side. The Thomas correlation with its published 1968 constants reaches 0.10, as recorded when the refitted baseline was set up. A Winland R35 index, the flow-oriented rock-typing estimate of Section 2.7 that collapses permeability and porosity into a single pore-throat radius and converts it through the Laplace relation, transfers at −0.88 as a parameter-free predictor and at −0.53 when a scaling is refitted under the same leave-one-source-out protocol. Both are far below the +0.46 of the refitted permeability-only law. The Winland result is the one anticipated previously; a single flow index cannot separate a porous low-permeability clay from a tight low-porosity evaporite, and forcing the two features into it destroys transferable signal instead of adding to it.

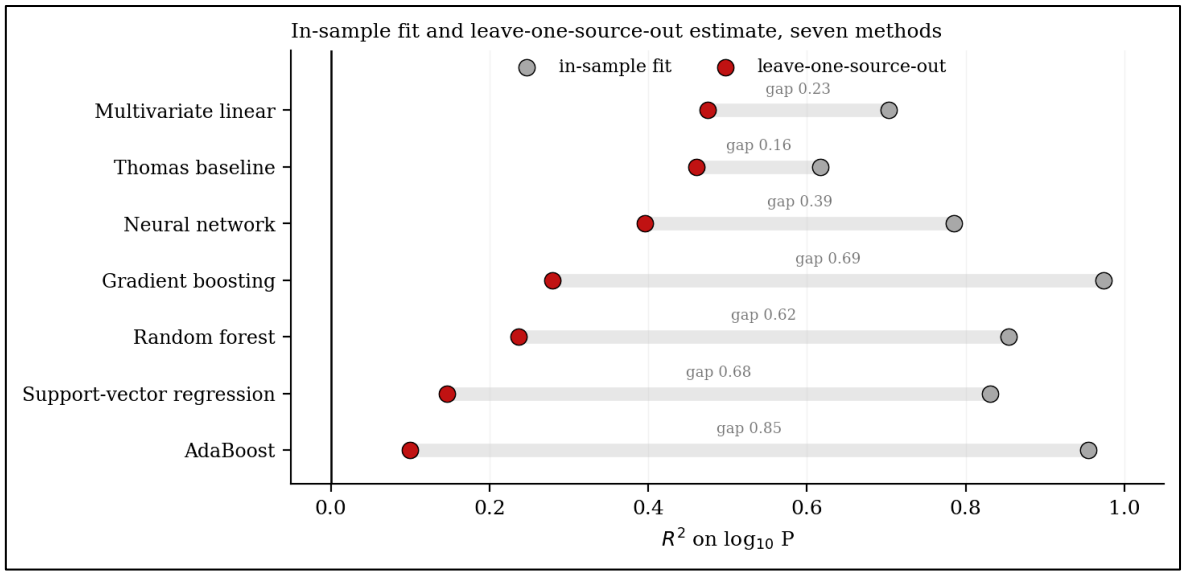


Figure 5.5 — The complexity stress-test. Each method is shown by its in-sample fit and its leave-one-source-out estimate, the span between them being the overfitting gap.

5.6 Feature contributions to transfer

Permutation importance measures how much accuracy falls when one input is randomly shuffled. It can be computed two ways, and on this dataset the two disagree. Computed on the fitted model, it gives 0.77 for porosity and 0.37 for log-permeability. Porosity appears to dominate.

Computed on the out-of-fold predictions, by shuffling a feature and re-running the whole leave-one-source-out procedure, it gives 0.23 for log-permeability and 0.03 for porosity over 200 permutations, with standard deviations of 0.05 and 0.03. Shuffling porosity barely changes what the model can predict for a study it has not seen. Recomputing the same importance under a random five-fold split, where records from one study fall on both sides of the partition, raises porosity's contribution to 0.086 against 0.37 for permeability: porosity matters about three times as much when the model is allowed to recognize the study, which is the between-study effect made visible.

The second number is the one relevant to the question asked here. In-sample importance describes how the fitted model uses a feature. Out-of-fold importance describes how much that feature contributes to predicting a study the model has not seen, which is the quantity this chapter reports throughout. The supported statement is therefore that permeability carries the transferable skill and porosity contributes little to it, and not that porosity dominates. That is consistent with Table 5.2, where adding porosity leaves the source-aware coefficient unchanged; its effect there is confined to the median error.

This agrees with most of the prior literature, which treats permeability as the dominant control and porosity as secondary. The in-sample importance, which would have suggested the opposite, is the figure a random-split validation would have reported, and it overstates porosity for a reason Chapter 3 anticipated. Section 3.7 showed porosity to be slightly the stronger of the two features by Spearman coefficient, and Section 3.11 showed that the group breaking the permeability ordering of breakthrough pressure is the one distinguished by its porosity; both are statements about the

pooled population. Section 3.9 showed that 86 % of the variance of porosity lies between studies. A correlation that is largely between studies is available to a model fitted on the whole population and is removed, at every fold, from a model that must predict a study it has not seen. The two importance figures differ because they measure the feature on either side of that removal.

Whether porosity acts as a proxy for study identity can be tested directly. Because most of porosity's variance lies between studies, holding a study out removes much of a porosity range as well as a set of rocks. A model leaning on porosity as a proxy for study identity must, at every fold, predict porosity values outside the range on which that proxy was learned, which increases its out-of-fold error. Leave-one-source-out tends to penalize a source-confounded feature, not reward it, and the collapse of porosity's importance from 0.77 in sample (random forest) to 0.03 out of fold (the leave-one-source-out linear model) is the size of that penalty. What survives is the permeability term, whose variance lies mostly within studies (Table 3.7). The argument would not hold under the random five-fold estimate, where same-study records sit on both sides of the partition, and that is one reason the two estimates are reported together.

This is where porosity parts company with confining pressure, discussed below, which carries a comparable between-study share of 87 per cent. A high between-study share does not mean the feature is constant inside a study, and for porosity it is not; six of the eight contributing studies report more than one porosity value, and the ranges inside a single study reach 0.114 to 0.400. Of the six studies that report an independent confining pressure, only one varies it at all. A model can therefore learn a porosity effect from contrasts inside a study, and the median-error gain of Table 5.2 is what that within-study contrast buys; any confining-pressure effect it learns has to come from contrasts between studies, which is exactly what source-aware validation cannot separate from study identity.

One limit on the claim remains. With eight sources the porosity ranges of the held-out studies are themselves few, so the test the design applies is a coarse one. The out-of-fold importance reported above removes the other limit, since it is measured on predictions made for studies absent from training.

5.7 Confining pressure as a feature

Chapter 3 set confining pressure aside because it is a study-level constant. The ablation tests that reasoning, and it confirms it in a way the raw scores alone would obscure.

Table 5.4 — Confining-pressure ablation, separating the feature from the population.

formulation	*n*	*sources*	R^2	*median error*
two features, no stress window (adopted)	*38*	*8*	*+0.47*	*× 1.97*
two features, records with an independent confining pressure	*24*	*6*	*−0.02*	*× 2.46*
three features, records with an independent confining pressure	*24*	*6*	*+0.36*	*× 1.56*

formulation	*n*	*sources*	R^2	*median error*
two features, 50–220 bar window	*27*	*7*	*+0.67*	*× 1.73*
three features, 50–220 bar window	*18*	*5*	*−0.11*	*× 2.45*

The nine records of Li, Gates & Hejazi (2025) are excluded from both three-feature rows. Their confining pressure was held 5 MPa above the injection pressure throughout the test, so at breakthrough it equals the breakthrough pressure plus 50 bar. Given to a model as an input, it is the target shifted by a constant. A three-feature model that includes those nine records is inflated by this leakage (it reaches +0.50 against +0.18 for the two-feature model on the same 33 records) and is not the comparison read here.

Holding the population fixed at the 24 records that report an independent confining pressure, the two-feature model reaches −0.02 and the three-feature model +0.36. Adding the variable improves the estimate by +0.38.

That improvement is not, on its own, evidence of a physical stress effect. In this compilation the variable is very nearly a source label. Six studies report an independent confining pressure. Five of them apply a single value to every one of their samples, and six of the seven distinct values in the population occur in one study only. Knowing the confining pressure of a record is therefore close to knowing which laboratory produced it. Eighty-seven per cent of its variance lies between studies.

Leave-one-source-out does not protect against a confound of this kind, and it is the main limitation of the design. The design prevents a model from recognizing a study it has already seen. It does not prevent a model from learning an association that exists between studies: that laboratories working at high stress happen to test tighter rocks. Such an association transfers to a withheld study without being physical, and it will raise the score exactly as a physical effect would.

The +0.38 therefore cannot be attributed to stress. It is what a near-unique study identifier is expected to produce, and nothing in this dataset can separate that from a stress effect. Section 3.10 gives the reason it cannot: the compilation records one stress value per study and not a stress series per sample, so the within-study contrast that would identify the physical effect does not exist.

Confining pressure is excluded because its contribution cannot be attributed. The data cost reinforces the decision without being its basis: requiring an independent confining pressure loses fourteen records and two sources, and requiring the stress window that makes them comparable loses six more, including every record above 297 bar. The −0.11 in the last row of Table 5.4 is a consequence of that reduction and not a measurement of the feature.

The two-feature score is itself sensitive to which records are present. On the 24-record population, which lacks the five ambient-condition records of Egermann et al. (2006) and the nine of Li, Gates & Hejazi (2025), the two-feature model reaches −0.02 against +0.47 on all 38. Removing only the five Egermann records, which include the lowest breakthrough pressure and the highest permeability in the population, lowers it to +0.18. That is a larger movement than any of the screening-sensitivity scenarios considered later in this chapter, and it shows that the transferable

score depends on the low-pressure, high-permeability end of the range being represented. The 50–220 bar window row moves in the other direction, to +0.67 on 27 records. That is the highest coefficient in the chapter, and it is not evidence for the window. The eleven records the window removes are the tightest rocks in the population, porosities of 0.001 to 0.030 and every breakthrough pressure above 297 bar; the 27 that remain span a narrower target, and Section 6.3 shows the coefficient of determination rising mechanically as the spread of the population falls. The window raises the score by discarding the hardest and most valuable rocks to predict, which is the opposite of what an adopted formulation should do. Which rocks are present moves the coefficient more than any feature does.

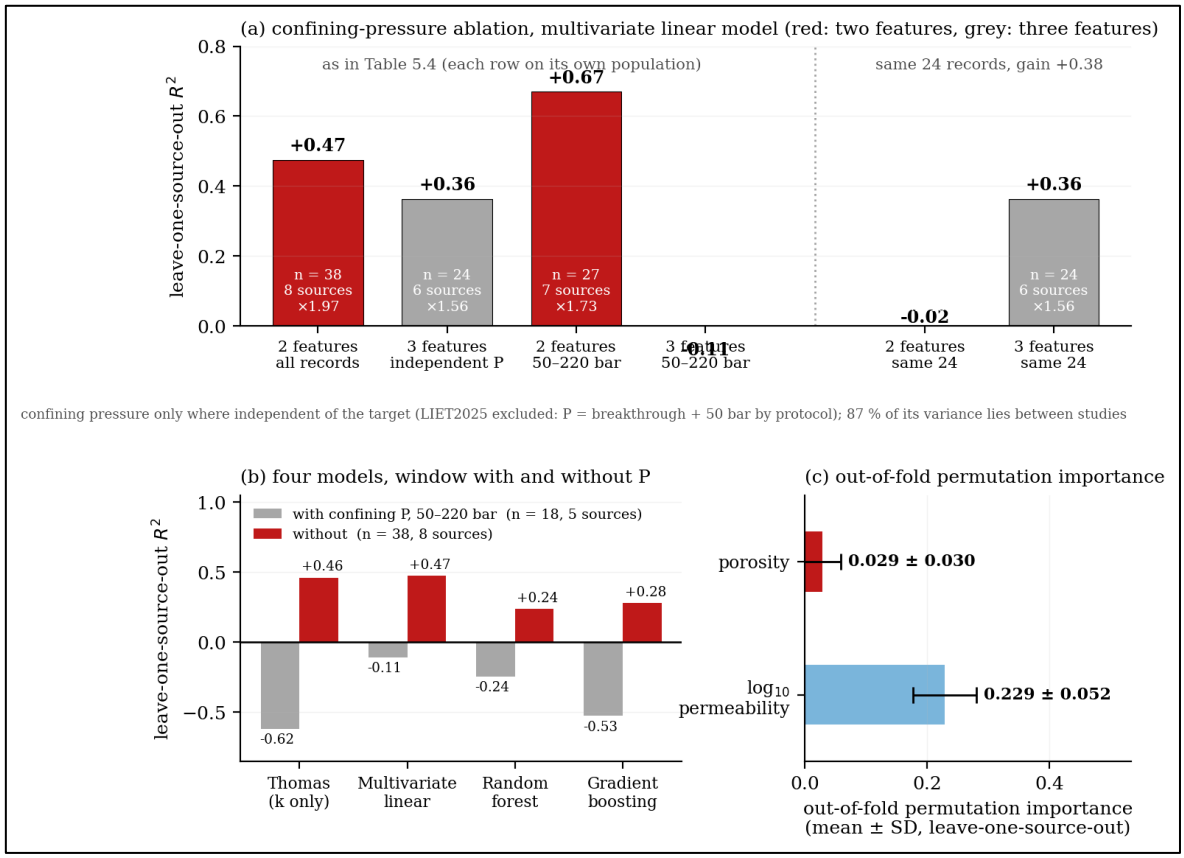


Figure 5.6 — Confining-pressure ablation, and how far the variable identifies its source study.

This does not call the physical effect into question. Higher effective stress narrows the conducting throats, and the literature demonstrates the effect directly on single plugs where the rock is held fixed and the stress varied. That experiment is the one this compilation does not contain.

5.8 Effect of dataset size

Whether the dataset limits the result of Section 5.4, and by how much, can be tested directly (Figure 5.7). The leave-one-source-out estimate is recomputed as sources are added to the training pool one at a time, from three studies to all eight, with the selection of studies resampled at each size.

The curve rises steeply from below zero at three sources to +0.44 at six, stays there at seven, and reaches +0.47 at eight. It has not flattened, but the last steps are small and no uncertainty is attached

to them, so the curve shows that the early sources matter and leaves open how much further sources would add. What it does not support is the opposite reading, that the method rather than the data sets the ceiling: nothing in the comparison of Section 5.5 points to the learner.

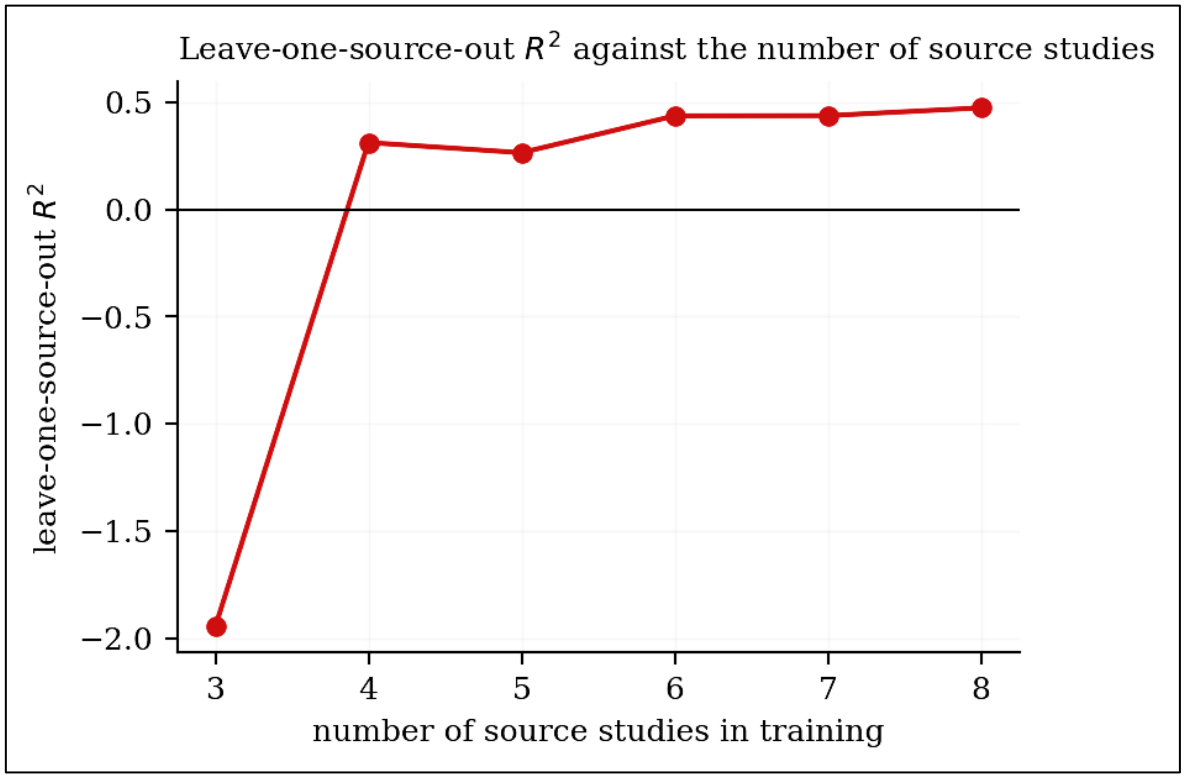


Figure 5.7 — Leave-one-source-out R² against the number of source studies used for training.

5.9 Snap-off pressure results

Snap-off pressure receives one section where breakthrough pressure received five, and the reason is established before any model is fitted. Chapter 2 showed the two quantities to be physically distinct: breakthrough is a drainage threshold set by the narrowest throat on the connected path, while snap-off is an imbibition instability set by the ratio of the pore body to the throat that drains it. Chapter 3 showed that the dataset records neither property directly: porosity shows no correlation with snap-off pressure, and permeability only a weak one that is not significant by rank.

The modelling that follows tests that expectation rather than explores it. It is run once, under the protocol used for breakthrough pressure, so the conclusion rests on a result and not on the argument alone.

All four models have negative leave-one-source-out scores.

Table 5.5 — Model comparison for snap-off pressure (n = 48, nine sources, 42 plugs). The interval is the 95 % record-level bootstrap interval.

model	in-sample fit	leave-one-source-out	95 % interval
Thomas baseline	*0.13*	*−0.37*	*−1.05 to −0.01*
Multivariate linear	*0.14*	*−0.44*	*−1.11 to −0.07*

model	*in-sample fit*	*leave-one-source-out*	*95 % interval*
Random forest	*0.60*	*−0.41*	*−1.08 to −0.07*
Gradient boosting	*0.88*	*−0.27*	*−1.00 to +0.13*

Every leave-one-source-out coefficient is negative, and a negative coefficient means the model predicts the withheld records worse than the pooled mean of the population would. The in-sample fits are not poor, gradient boosting reaches 0.88, which shows the failure is not an inability to describe the data but an inability to carry any description to a new rock. The multivariate model, which was the best performer on breakthrough pressure, is the worst here, and that is the expected consequence of adding a feature that carries no signal for this target.

The outcome matches the expectation set out above: the property that governs snap-off is not among the inputs, and no model can recover from bulk transport measurements a geometric ratio they do not encode. Snap-off pressure is therefore not carried forward as a predictive target, and Chapter 6 returns to how it should be treated instead.

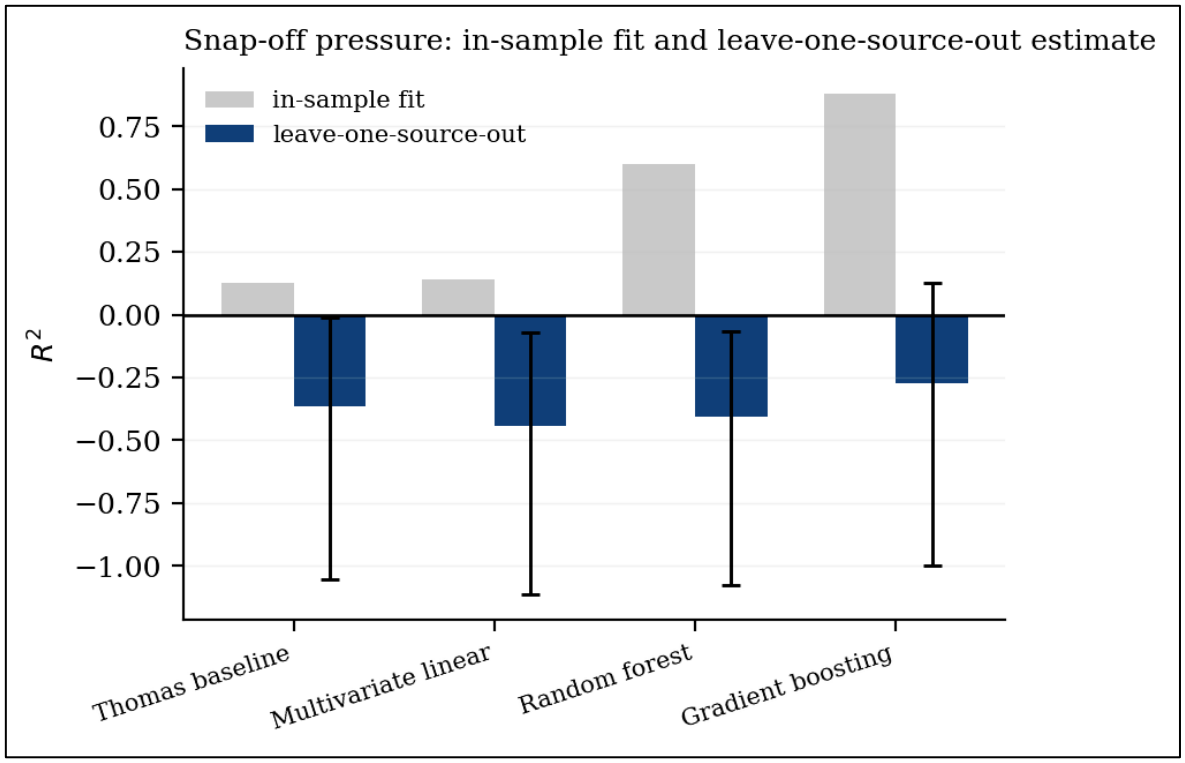


Figure 5.8 — Snap-off pressure: in-sample fit and leave-one-source-out estimate for each model.

5.10 Sensitivity to screening decisions

Chapter 3 removed nearly two thirds of the merged compilation, and every removal rested on a judgement made by the author. Whether the result above is a property of the data or a product of those judgements can be tested. The screening comprises seven exclusion steps after duplicate removal (Table 3.2). Each was reversed in turn, the excluded records restored, and the whole modelling procedure re-run unchanged. For the compiled-source step only the literature compilation

can be restored, since the field-scale review carries no petrophysics. A final scenario applies no screening beyond duplicate removal.

Table 5.6 — Transferable accuracy under alternative screening decisions (multivariate linear, leave-one-source-out).

scenario	n	sources	R^2	median error
adopted screening	*38*	*8*	*+0.47*	*× 1.97*
keep the compiled source	*38*	*8*	*+0.47*	*× 1.97*
keep the placeholder block	*49*	*9*	*+0.46*	*× 2.14*
keep the unattributable records	*48*	*12*	*+0.47*	*× 2.09*
keep the imputed-feature records	*46*	*9*	*+0.52*	*× 1.92*
keep the figure-derived records	*44*	*9*	*+0.42*	*× 2.18*
keep the crack-dominated records	*38*	*8*	*+0.47*	*× 1.97*
keep the out-of-scope reservoir rock	*40*	*9*	*+0.52*	*× 1.97*
no screening at all	*75*	*16*	*+0.52*	*× 1.93*

Two reversals leave the population unchanged. The compiled source contributes only mercury-injection values, which the breakthrough population excludes by method, and the crack-dominated records are snap-off measurements. Restored unattributable records are grouped for validation under the source label they carry in the dataset. Section 3.4 argues that such a grouping has no basis, so that row is shown for completeness and not as a valid source-aware estimate.

The estimate stays between 0.42 and 0.52 in every scenario, including the one with no screening. The headline figure of 0.47 sits inside that range. No single exclusion carries the result, and the result is not an artifact of the screening.

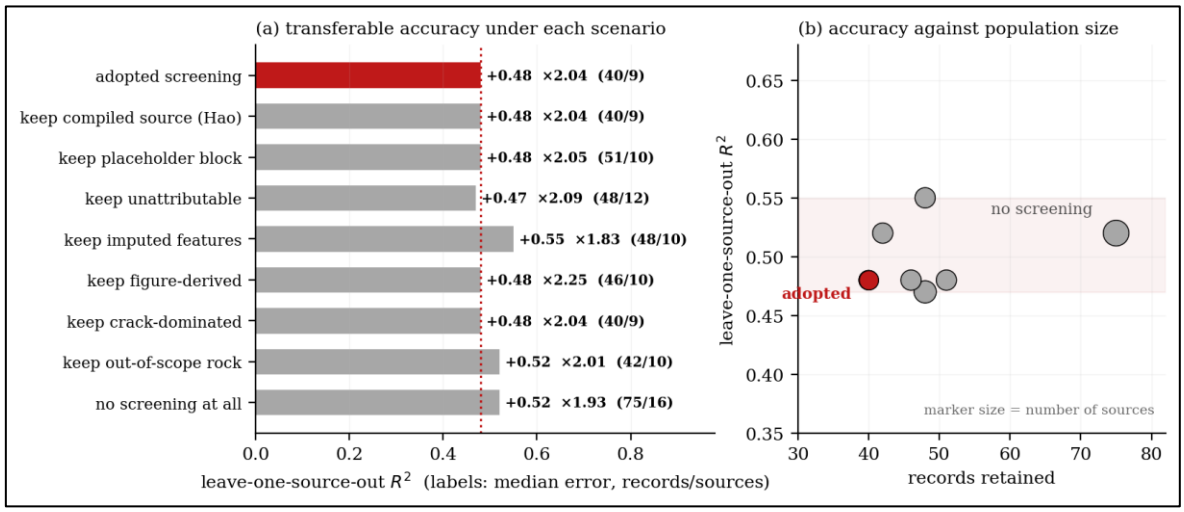


Figure 5.9 — Transferable accuracy under alternative screening decisions.

The screening sensitivity supports two conclusions.

The first is that the result is stable under the screening. The central claim of this thesis does not depend on the author's judgement about which records to remove, and a reader who disagreed with every exclusion would reach the same conclusion. Stability under screening is not stability in general; as shown earlier in this chapter, removing the five ambient-condition records of Egermann et al. (2006), which carry the lowest breakthrough pressure and the highest permeability in the population, lowers the estimate to +0.18. The estimate depends on which rocks are present far more than on which records are screened out.

The second is a caution about what the verification achieved. If the accuracy is the same either way, the verification of Chapter 3 did not improve the number. What it did was make the number mean something. A source-aware estimate requires that every record be assignable to a source, which the unattributable block prevents. A model trained on records whose experimental conditions are placeholder values is not learning the physics it appears to learn, whatever it scores. And the record counts, the source counts and the fold structure that Chapter 6 uses to bound the claim are only meaningful if the provenance behind them is real.

5.11 Summary of results

Four models were fitted to breakthrough pressure and compared under source-aware validation. The comparison gives a modest result. A source-validated model predicts the breakthrough pressure of an unseen study with a coefficient of determination of 0.47, a root-mean-square error of 0.50 log units (a factor of 3.2) and a median error of a factor of two. Its record-level interval excludes zero, as does that of the refitted permeability-only correlation, though a source-level bootstrap leaves every interval spanning zero.

That model is the multivariate linear regression, not an ensemble. Both tree methods transfer worse than either straight line while fitting their training data better. Admitting porosity as a second input does not raise the transferable coefficient above that of the permeability-only law; it lowers the median error from a factor of 2.6 to 2.0, and that is the whole of its contribution. Accuracy measured without regard to provenance would have been roughly two fifths higher for the linear model and two and a half to three times higher for the ensembles, and would have been misleading.

Confining pressure is excluded because it is very nearly a source label in this compilation: of the six studies reporting an independent value, five apply a single value to all their samples, and six of the seven distinct values occur in one study only. Adding it raises the estimate on a fixed population, but leave-one-source-out cannot separate that gain from study identity, so the gain is uninterpretable. The transferable signal rests on permeability; measured out of fold, porosity contributes little to it. The learning curve still rises at eight sources, though its last steps are small, so the number of studies remains the more likely constraint.

The result does not depend strongly on the screening. Reversing each of the seven exclusions in turn, and applying no screening at all, moves the estimate only between 0.42 and 0.52. What the verification of Chapter 3 buys is not accuracy but interpretability: a source-aware estimate is only meaningful if every record can be assigned to a source.

Snap-off pressure, modelled under the same protocol, fails by every model and every measure, confirming the physical expectation set out in Chapters 2 and 3. Chapter 6 takes up what these results mean for the estimation of sealing capacity in practice.

Chapter 6 — Discussion and Conclusions

The thesis set out to predict caprock sealing pressure from petrophysical properties and to test that prediction against the correlation the field already uses. Chapters 3 to 5 supplied the parts of an answer; a verified dataset, a chart that needs no computer, and a model comparison judged by transfer to unseen studies. This chapter draws them together. Its subject is the boundary of predictability; which sealing quantity is predictable, from which measurements, to what accuracy, and why the boundary sits where it does.

Three findings frame the discussion, and each differs from what the published literature on this problem reports. The transferable accuracy of a data-driven model is well below the values reported in published studies, and the difference is traced to how those studies validate. No model transfers better than a refitted permeability-only correlation, which the framework therefore adopts as its predictive equation. And the coefficient of determination alone is a poor measure of accuracy for this problem.

6.1 Summary of the modelling results

Breakthrough pressure is predictable from permeability and porosity with modest accuracy, and only under record-level resampling can that accuracy be distinguished from zero. The multivariate linear regression reached a leave-one-source-out coefficient of determination of 0.47, with a record-level bootstrap interval of +0.15 to +0.68; the refitted Thomas correlation reached 0.46 with an interval of +0.12 to +0.68. Those two intervals exclude zero and the two ensembles do not.

Resampling whole studies widens the multivariate interval from −0.37 to +0.71, so the claim is that the two linear models are the best supported of the four and not that either is demonstrably better than predicting the pooled mean. The multivariate model's median prediction error is a factor of 2.0. The random forest reached 0.24 and gradient boosting 0.28, with intervals straddling zero under either bootstrap.

The record-level interval on the multivariate estimate spans 0.53, the source-level interval 1.09, and the four models intervals overlap heavily, the Thomas baseline estimate lies within 0.01 of the multivariate model and its upper bound coincides with it. The comparison therefore identifies the best supported models, not one that is demonstrably superior to the classical correlation. Section 6.9 returns to how small the practical gain is.

Two secondary results are better supported by the data than the model ranking, whose intervals overlap. Model flexibility does not improve transfer; both tree ensembles transfer worse than either straight line while fitting their training data better, gradient boosting reaching 0.97 in sample and 0.27 out of it. And confining pressure is excluded because its contribution cannot be attributed, not because it is absent. Adding it raises the coefficient from −0.02 to +0.36 on a fixed population of 24 records. In this compilation, however, five of the six studies that report an independent confining pressure apply a single value to every sample. Six of the seven distinct values occur in one study only, so the variable is very nearly a source label.

Source-aware validation prevents the model from recognizing a study it has already seen; it does not prevent it from exploiting an association that holds between studies. The gain therefore cannot be attributed to geomechanical stresses.

6.2 Robustness to the removal of a source

To evaluate the model robustness independence of individual publications, each of the contributing studies was sequentially withheld, recomputing the leave-one-source-out validation on the remaining seven.

Table 6.1a — Leave-one-source-out accuracy with each source withheld in turn.

Source withheld	Records removed	Records left	R^2	Median error
none (all eight)	—	38	+0.47	× 1.97
Rezaeyan et al. (2015)	6	32	+0.74	× 1.87
Klimkowski & Smulski (2012)	3	35	+0.50	× 1.88
Amann et al. (2011)	1	37	+0.47	× 2.01
Ito et al. (2011)	3	35	+0.45	× 2.03
Li et al. (2005)	9	29	+0.44	× 2.10
Kawaura et al. (2013)	2	36	+0.43	× 2.07
Li, Gates & Hejazi (2025)	9	29	+0.27	× 3.05
Egermann et al. (2006)	5	33	+0.18	× 2.14

The coefficient stays positive under every single-source removal, but the spread is wide, from +0.18 to +0.74. Two removals stand out. Withholding Rezaeyan et al. (2015) raises the coefficient to +0.74; those six anhydrite and shale records are the tightest and highest-pressure rocks in the population and the least well predicted, so removing them narrows the target range and raises R^2.

At the other end, withholding Egermann et al. (2006) drops it to +0.18 and withholding Li, Gates & Hejazi (2025) to +0.27. The Egermann records are the five ambient-condition, low-pressure, high-permeability points that anchor the bottom of the range; the Li, Gates & Hejazi records supply three of the highest breakthrough pressures below 300 bar together with the only hydrogen measurements. Both anchor an end of the target range, the estimate depends on having both ends represented

This is a change from what a smaller and less thoroughly verified version of this dataset showed, where removing the largest source collapsed every model below zero. The verification of Chapter 3 removed a twenty-five-record block whose values were traceable to three different studies but whose features and conditions matched none of them; that block was the largest single contributor to the earlier breakthrough population, and a model relying on it was relying on records that should not have been present. The coefficient no longer falls below zero when a source is withheld, but

with eight sources it still depends materially on which of them are present, and in particular on the two that anchor the ends of the pressure range.

This indicates that previous estimates were biased by an artificial subset within the compilation, and, that the discrepancy stems not from a methodological flaw in the earlier analysis, but from its reliance on a dataset distorted by a major data artifact. Consequently, the revised findings offer more tempered claims while explicitly detailing their residual sensitivity to individual data sources.

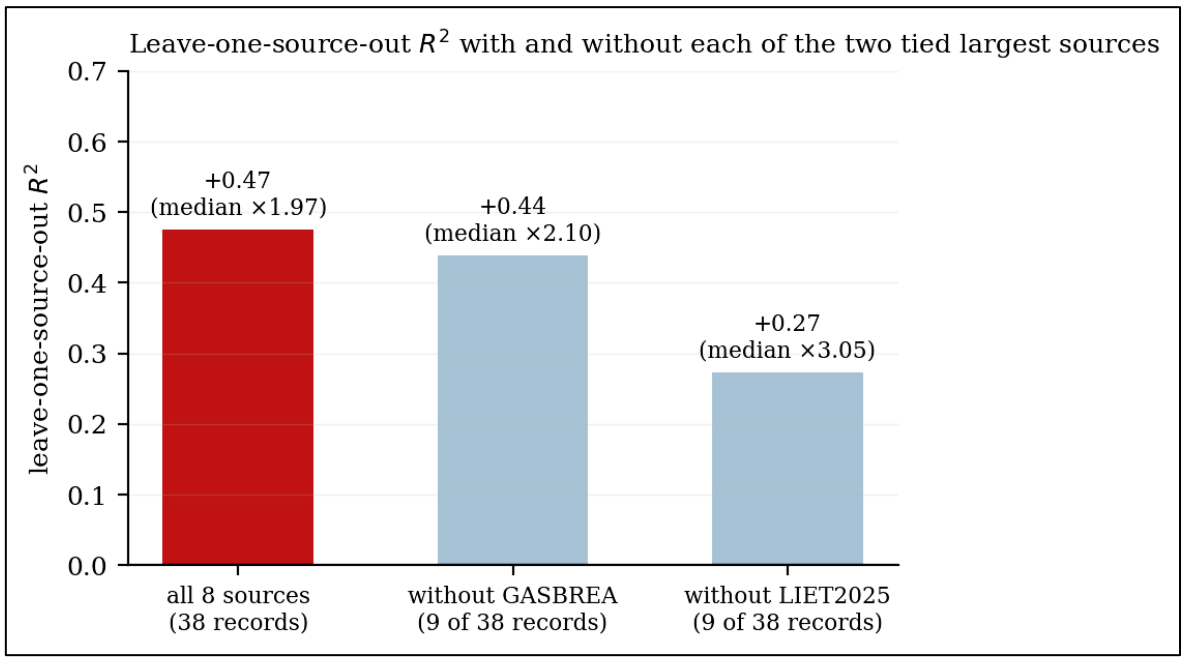


Figure 6.1 — Leave-one-source-out accuracy with all eight sources and with each source withheld in turn (Table 6.1a).

6.3 Choice of accuracy measure

Under the operational conventions adopted in this study, formations exhibiting breakthrough pressures below 10 bar are categorized as non-sealing, while those between 10 and 20–30 bar represent poor seals. These low-pressure records were deliberately retained in the dataset to train and evaluate models across the complete spectrum of sealing behavior. Consequently, assessing model skill specifically on samples that meet practical seal criteria provides a more operationally relevant measure of predictive accuracy.

Table 6.1 — Model performance by breakthrough-pressure threshold (multivariate linear, leave-one-source-out).

records retained	n	sources	R^2	median error	spread of target (SD, log units)
all	38	8	+0.47	× 1.97	0.71
≥ 10 bar (is a caprock)	29	7	−0.04	× 1.96	0.49
≥ 20 bar (adequate seal)	26	5	−0.41	× 1.91	0.43
≥ 30 bar (good seal)	20	5	−1.27	× 1.97	0.36

The coefficient of determination falls sharply. The median prediction error does not; it stays close to a factor of two across every subset.

The apparent contradiction arises from the measure. The coefficient of determination expresses the fraction of the variance of the target that a model explains, so it depends on how much variance there is to explain. Restricting the population to good seals removes half the spread. The standard deviation of the target falls from 0.71 to 0.36 log units, and the number of contributing studies falls from eight to five, so the estimate is also averaged over fewer and larger folds. And a model can predict every value within a factor of two while explaining little of what little variance remains. On a narrow range, a low coefficient of determination is not an inaccurate model.

For an engineering purpose, the more useful quantity is the accuracy of a prediction, expressed as a factor on the measured value. The fraction of variance explained across a heterogeneous population is less suited to that purpose. This work therefore uses the coefficient of determination to compare models on one population, and reports the median error beside it wherever populations differ, since two coefficients computed on populations of different spread are not comparable. Everywhere the accuracy of an estimate is at issue, the error is reported as a factor.

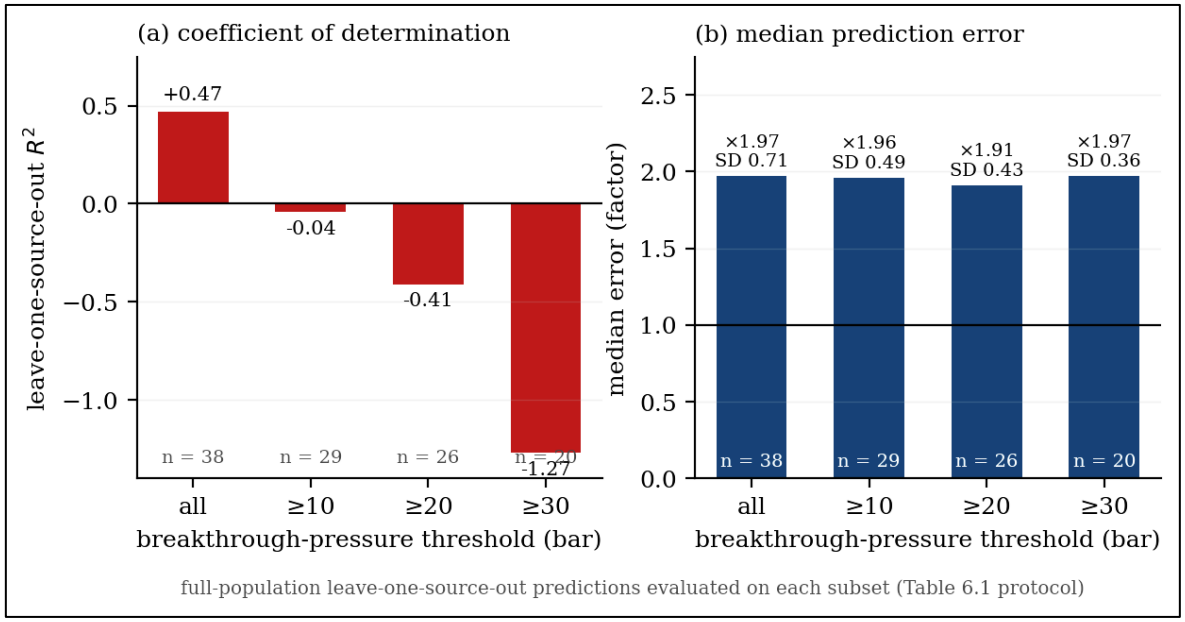


Figure 6.2 — Restricting the population to better seals.

6.4 Accuracy by seal-quality class

The three petrophysical classes of Chapter 3 provide the comparison that lithology labels cannot. Predictions taken from the leave-one-source-out folds, so that every one is made on an unseen study, fall within a median factor of 1.66 for the ultra-tight class, 1.70 for the clay-rich class and 2.40 for the moderate class.

The ultra-tight class, whose median breakthrough pressure is 200 bar, contains the genuine caprocks, and the model predicts them within a factor of 1.7. The clay-rich class, at the opposite end of the porosity range, is predicted as well, though on six records only. The moderate class is

the least accurate, and it is also the most permeable of the three, at a median breakthrough pressure of 27 bar.

The model is therefore not weaker on good seals, as the falling coefficient of determination on higher-threshold subsets might suggest. It is weakest on the most permeable class, whose median breakthrough pressure of 27 bar falls in the poor-seal band established for the threshold analysis above.

Table 6.2 — Prediction accuracy by seal-quality class, from leave-one-source-out folds.

seal-quality class	***n***	***median measured P_b (bar)***	***median error (factor)***	***within a factor of two***
ultra-tight	*12*	*200.0*	*× 1.66*	*75 %*
clay-rich	*6*	*9.5*	*× 1.70*	*67 %*
moderate	*20*	*27.0*	*× 2.40*	*35 %*
all	*38*	*41.5*	*× 1.97*	*53 %*

A stronger test of generalization is to withhold a whole seal-quality class, not a study. Training the model on two of the three classes and predicting the third, in turn, gives an overall coefficient of 0.62 and a median error of a factor of 1.90; the ultra-tight class is predicted within × 1.61 when the model has seen only the clay-rich and moderate classes, the clay-rich within × 1.89, and the moderate within × 2.07.

Predictive transfer to an unseen rock class is therefore at least as effective as transfer to an unseen study, yielding a higher coefficient of determination (R^2 = 0.62 vs. 0.47). This occurs because the seal-quality classes occupy contiguous domains within the porosity–permeability plane, allowing the linear relationship to smoothly extrapolate across class boundaries. This is the closest the data come to testing generalization beyond the studied rocks, and it does not break down.

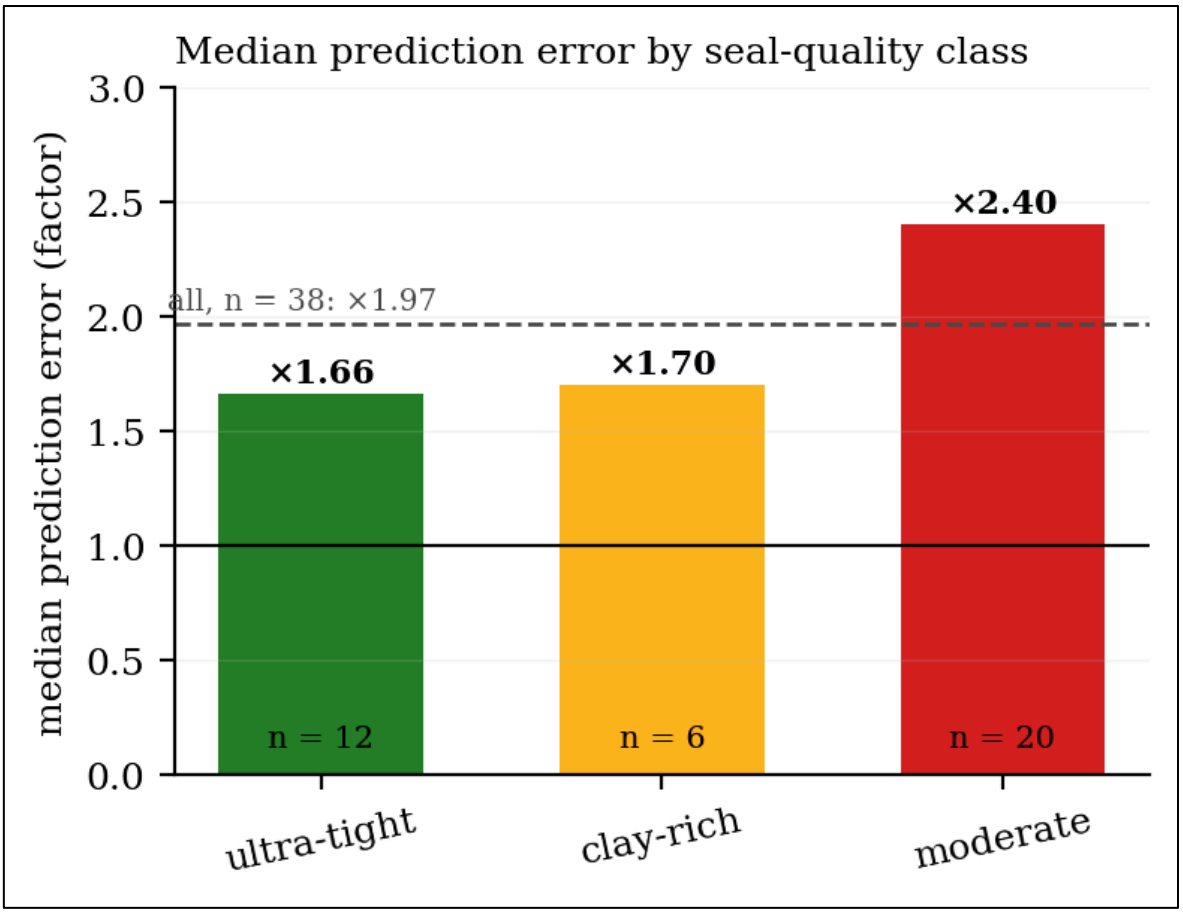


Figure 6.3 — Median prediction error by seal-quality class.

6.5 Estimation from the charts alone

Chapter 4 supplied a second route to the same quantity. Chart BP-1 expresses breakthrough pressure as a function of permeability and porosity, fitted to the same 38 records with a coefficient of determination of 0.70 in sample and a scatter band of a factor of 2.4. Chart BP-2 converts a mercury-injection entry pressure through the ratio of the capillary systems, with factors of 0.16 to 0.19 for the storage fluids at a zero-contact angle.

The mercury conversion was checked against five same-plug comparisons in the compilation, of which three agree within a factor of two, one is over-estimated by a factor of ten and one under-estimated by a factor of almost four. Two further studies show the conversion failing in opposite directions on single plugs. The conversion is then an order-of-magnitude estimate, useful for a first assessment from a test that takes hours instead of weeks, and not a substitute for a flow-through measurement.

6.6 Chart and model compared

The framework offers two routes to breakthrough pressure, and the choice of which relationship to adopt as its predictive equation is settled by the comparison of Chapter 5. The equation whose transfer has been measured, and which nothing more elaborate improves on, is the refitted

permeability-only law of Equation 5.2. It is adopted as the framework's predictive equation for three reasons.

First, on the rocks that would actually be assessed as caprocks, at 10 bar and above, it has the smaller median error of the two linear models (× 1.93 against × 1.96 at 10 bar and above, × 1.86 against × 1.91 at 20 bar and above, Table 6.5).

Second, its transfer is distinguishable from zero under the record-level bootstrap, though, as for every model here, that distinction does not survive the plug- or source-level resampling; the two-feature model's interval also excludes zero at plug level.

Third, admitting porosity, the only refinement the data offer, does not raise the transferable coefficient above it (0.47 vs. 0.46). A refitted single-parameter permeability correlation, validated source by source, is what the data support as the working estimator.

Porosity is retained in Chart BP-1. Where both properties are known, the chart's two-feature reading lowers the median error across the full population from a factor of 2.6 to 2.0 and carries an in-sample band of a factor of 2.4.

Because these values describe fundamentally different performance measures, they must be distinguished, the factor of 2.4 quantifies in-sample scatter around the regression fit, while the transfer statement reflects out-of-sample generalization across unseen datasets. Put on the same out-of-fold predictions, 53 % fall within a factor of two of the measured value, 63 % within the chart's band of 2.4, and 68 % within the RMSE factor of 3.2. Thus, treating the chart's empirical uncertainty band as a two-thirds confidence threshold for new sites slightly understates the out-of-sample variance, given that 63% of out-of-fold predictions fall within this range

The practical output is therefore a tiered framework in which the tiers are set by the measurements available. Table 6.3 states what each rests on and what accuracy it can support. Chart BP-1 is best read as one fitted relationship carrying two uncertainty statements rather than as two competing equations: where both permeability and porosity are known it gives the two-feature reading with an in-sample scatter band of a factor of 2.4, and its transfer to an unseen study is the leave-one-source-out estimate of Chapter 5; the adopted predictive equation, Equation 5.2, is its permeability-only reading, used where porosity is unavailable or where only the transfer-measured coefficient is wanted.

Table 6.3 — The tiered estimation framework: route, evidence and stated accuracy.

tier	***measurements available***	***route***	***accuracy statement***	***basis***
1	*mercury-injection entry pressure only*	*Chart BP-2 (Equation 4.2)*	*order of magnitude: three of five same-plug checks within a factor of two, the other two out by ×10 and ×0.26*	*Table 4.3*

tier	measurements available	route	accuracy statement	basis
2	*permeability only (adopted equation)*	*Equation 5.2, the refitted permeability-only law*	*on unseen studies: R^2 0.46, median error ×2.6 over the full range and ×1.86 above 20 bar, RMSE ×3.2; record-level interval excludes zero, source-level interval does not*	*Tables 5.2, 6.5*
3	*permeability and porosity*	*Chart BP-1 (Equation 4.1) as a refinement*	*in-sample band ×2.4, covering 63 % of out-of-fold predictions; porosity lowers the full-range median error to ×2.0 but does not raise the transferable coefficient*	*Section 4.1; Section 6.6*
-	*snap-off pressure, any tier*	*not predicted; bounded by the class median*	*class median with about half the measured values below it and fewer than half within a factor of two*	*Table 6.4*

Snap-off pressure enters no tier as a prediction; it is described by the range measured within the sample's seal-quality class (Section 6.7).

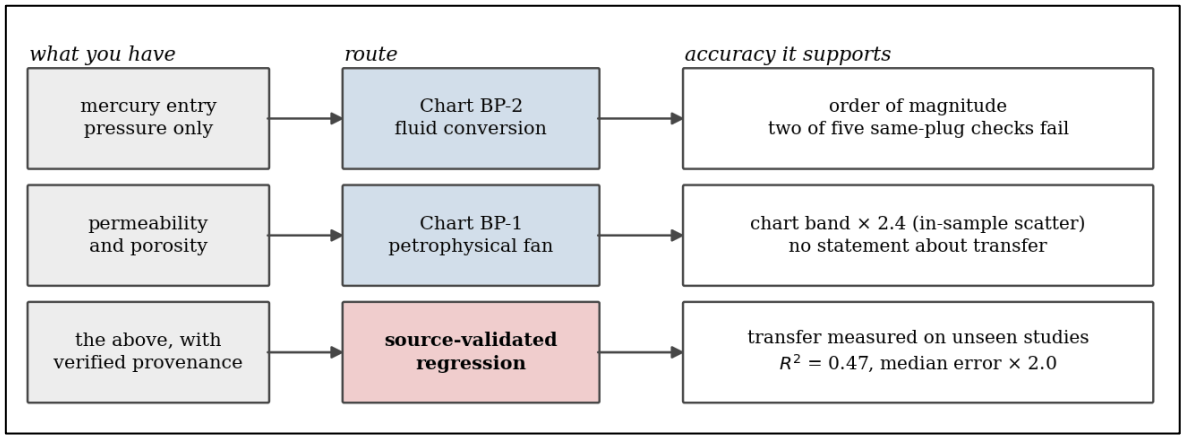


Figure 6.4 — The tiered estimation framework. Each route is matched to the measurements available and to the accuracy it can support; snap-off pressure is bounded by class and not predicted.

6.7 Limits of predictability

The framework stops at breakthrough pressure. Snap-off pressure failed every model under the same protocol, with leave-one-source-out coefficients between −0.27 and −0.44 and intervals almost entirely below zero, while fitting the training data as well as 0.88.

This predictive limitation is more likely a consequence of physical information lack in the input features and not modeling techniques deficiency. Breakthrough is a drainage threshold set by the narrowest throat on the connected path, a quantity that permeability and porosity jointly constrain. Snap-off is an imbibition instability governed by the ratio of the pore body to the throat that drains it. That ratio is a geometric property which bulk transport measurements do not record, and the fluid–rock properties that also bear on it are absent from the compilation at the sample level. A second possibility, raised in Section 2.4, is that the residual quantity is not a material property at all,

in which case no rock descriptor would predict it. A third possibility lies in the data itself. The snap-off population is more concentrated than the breakthrough one, eighteen of its forty-eight records come from a single source (Hildenbrand et al., 2004), and five are repeat carbon-dioxide exposures on two plugs whose sealing capacity the source itself reports degrading between runs (Wollenweber et al., 2010). This within-study scatter is therefore partly attributable to target non-stationarity across sequential measurements, which predictive models treating the response as a static property cannot resolve. None of the three is tested here, since the aspect ratio is not reported in any source and the population is too small to separate them.

Snap-off pressure is therefore better bounded than predicted. The verified data gives a pooled ratio of snap-off to breakthrough median of 0.19, just below the 20 to 50 % range the literature reports, but on the thirteen plugs that carry both measurements the ratio runs from 0.22 to 0.92 (Section 3.8), so no single fraction of breakthrough pressure describes snap-off.

A class-specific bound is the simplest alternative, and its construction is set out here.

Table 6.4 — Class-specific snap-off bound, and the spread of measured snap-off about it.

seal-quality class	***implied bound (bar)***	***measured values at or above the bound***	***measured values within a factor of two of the bound***	***n***
clay-rich	*7.0*	*53 %*	*37 %*	*19*
moderate	*6.3*	*55 %*	*45 %*	*11*
ultra-tight	*6.9*	*50 %*	*44 %*	*18*

This implied bound evaluates directly to the class median snap-off pressure, rendering its value independent of breakthrough pressure. Consequently, the resulting snap-off bounds across the three seal-quality classes vary by under 10% (6.3–7.0 bar), even though their corresponding breakthrough medians span a factor of twenty. By definition of a median estimator, approximately 50% of measured values lie at or above the bound, though fewer than half reside within a factor-of-two margin.

This bound represents a baseline description of within-class spread, yet, it is not conservative, as roughly half of measured snap-off pressures in each class fall below it. In practice, it carries the same order-of-magnitude status as the mercury conversion in Chapter 4, serving as a practical, preliminary estimate with explicit uncertainty bounds for containment assessments when predictive models are unavailable

6.8 Comparison with published work

Published data-driven studies report much higher accuracy, and this work approaches that magnitude when it validates in the same way. One compilation study reports a neural network at R^2=0.99 on a random split of a pool drawn from thirteen source studies; the same study's own pooled permeability correlation reaches only 0.49. In the present work the random five-fold estimate for the multivariate model is 0.67 and the in-sample fit of gradient boosting is 0.97, against a leave-

one-source-out estimate of 0.47. The difference lies mainly in what the validation protocol. A random split places records from the same study on both sides of the partition, so a model can score well by recognizing a study it has already partly seen. The contribution of this thesis is the demonstration, on a source-attributed and provenance-verified compilation, that transferable accuracy for this problem is substantially below what random-split validation reports.

The permeability-only correlation transfers as well as the two-feature model on the coefficient of determination, and its transfer can be distinguished from zero only under the weaker bootstrap. The refitted Thomas relation reaches 0.46 with a record-level interval from +0.12 to +0.68, widening to −0.71 to +0.78 when whole studies are resampled. The first interval excludes zero and the second does not, so on this dataset the correlation can be shown to predict an unseen study better than the pooled mean only if records within a study are treated as independent. These results complement existing literature by establishing its empirical boundaries. The classical correlation was historically intended as a narrow, first-order approximation, and its performance limit emerges here primarily because its transferability had not yet been benchmarked against unseen source studies

Rocca (2021), compiling literature threshold and residual data for nitrogen and testing the Thomas correlation against them, concluded that it is a reliable predictor of the order of magnitude. The present work extends that in four directions. It covers four fluids and not one. It separates breakthrough from snap-off as distinct targets. It quantifies transferable skill in place of assessing agreement visually. And it shows that a data-driven petrophysical grouping separates the target, which the lithology labels do not do at this sample size (Table 3.4), although the groups are built on the two variables that predict it.

Evaluated out-of-fold, porosity contributes minimally to transferable predictive skill (permutation importance of 0.03 versus 0.23 for log-permeability), consistent with established literature identifying permeability as the primary petrophysical control.

Within the population, porosity is the stronger of the two features by rank correlation and it lowers the median error of the transferred prediction from a factor of 2.6 to 2.0, exceeding expectations established by conventional petrophysical models. The two parts are reconciled by Section 3.9; 86 % of the variance of porosity lies between studies, so most of what porosity explains in a pooled fit is not available to a model asked to predict a study it has not seen. Two studies report the opposite sign for porosity. A possible explanation is that the sign depends on whether a sample suite spans a compaction trend or a lithofacies contrast, which would itself explain why a single pooled correlation transfers poorly.

This inability to predict snap-off pressure aligns with theoretical description of the pore-scale literature. If snap-off is governed by the pore-body-to-throat aspect ratio, its resistance to prediction from bulk porosity and permeability is the expected consequence.

6.9 Gain over the Thomas correlation

A comparison on equal terms is needed, because the case for a data-driven approach depends on it. Both were validated identically; Table 6.5 takes the median error of each from the same leave-one-source-out predictions as Table 6.1, evaluated on the same threshold subsets.

Table 6.5 — Median prediction error, Thomas baseline against the multivariate model, leave-one-source-out.

Records retained	n	Thomas baseline	Multivariate linear
all records	*38*	*× 2.60*	*× 1.97*
≥ 10 bar	*29*	*× 1.93*	*× 1.96*
≥ 20 bar	*26*	*× 1.86*	*× 1.91*
≥ 30 bar	*20*	*× 1.86*	*× 1.97*

The error-reduction benefit of the multivariate model is confined to the broader, full-range dataset and does not extend to high-quality seals. Although adding porosity cuts the overall median prediction error by 25%, the single-parameter permeability model performs slightly better on samples with breakthrough pressures at or above 10 bar (×1.93 vs. ×1.96) and 20 bar (×1.86 vs. ×1.91). This demonstrates that the porosity term primarily corrects for over-predictions on non-sealing formations (<10 bar). On the data available, the refitted single-parameter correlation improves it by a margin confined to the low end of the range.

The advantage of adopting the refitted baseline is realized not relative to the two-feature linear regression, which yields comparable transferable accuracy, but relative to conventional implementations of the Thomas relation. With its published 1968 constants the Thomas relation reaches only 0.10 and a median error of a factor of 3.2 on this population; refitting the two constants to the data raises it to 0.46 and ×2.6, a larger gain than porosity then adds. This is why porosity was kept only in the chart; the porosity term contributes on the weak rocks and in the full-range median error, not on the caprocks, and the correlation carries the smaller error where a containment decision is actually made. Both linear models have now had their transfer to an unseen study measured, and both record-level intervals exclude zero while neither source-level interval does; a practitioner using either has an estimate whose transferability has been quantified, and quantified as modest, which is more than the correlation carried before.

As illustrated in Figure 6.5b, both models exhibit a systematic bias shift across the pressure spectrum, over-predicting low-integrity formations while under-predicting high-capacity seals. This sign reversal reflects a classic regression to the mean, characteristic of models fitted to small, heterogeneous datasets.

In containment assessments, while underestimating caprock performance provides an inherent safety margin, overestimating low-integrity seals introduces non-conservative operational risk, making this optimism the primary decision-critical failure mode.

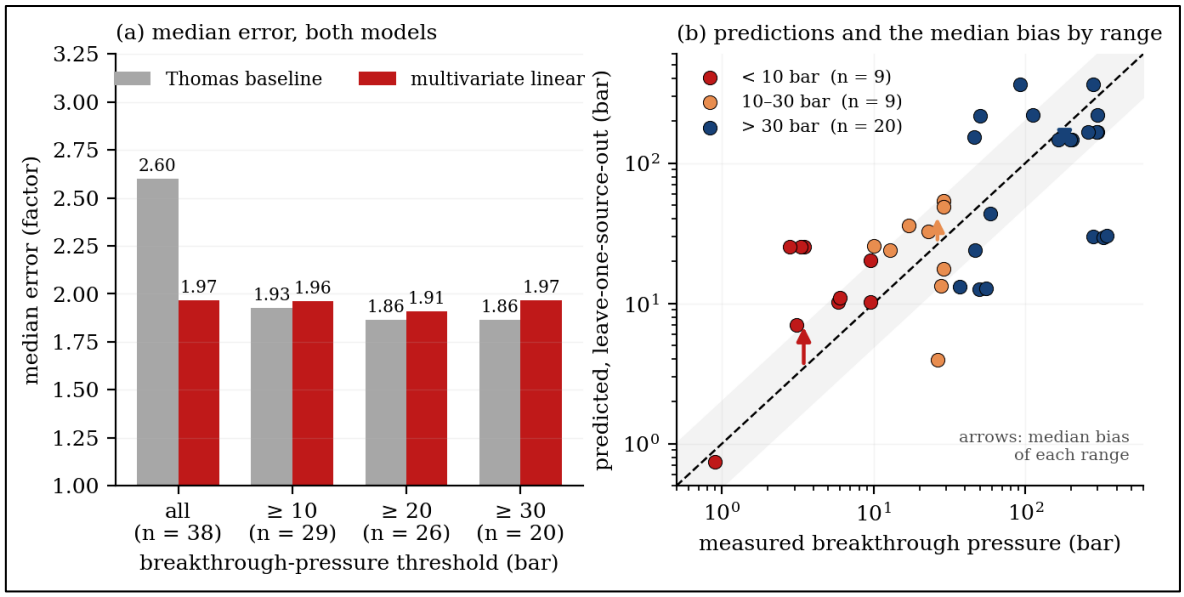


Figure 6.5 — Median error of both models across thresholds, and the residual bias by pressure range.

6.10 Limitations

The dataset is small and unevenly sourced; 38 breakthrough records from eight studies. Source-aware validation is applied at publication level, and the publications are not fully independent. Ito et al. (2011) and Kawaura et al. (2013) share several authors and a common measurement suite, and Hildenbrand et al. (2002) and (2004) share authors and an apparatus, so the effective number of independent groups is smaller than eight and the source-level intervals, wide as they are, are if anything optimistic. The learning curve had not flattened when the data were exhausted, although its last steps were small, so the number of independent studies is the more likely a constraint not a choice of method.

The model is most accurate on the ultra-tight class and least accurate on the moderate class, as shown earlier in this chapter, and Section 5.7 shows the coefficient falling to +0.18 when the five ambient-condition, high-permeability records of Egermann et al. (2006) are removed. Read together, these say that a large part of the transferable skill is the model separating a weak, permeable rock from a tight seal. The coefficient of determination rewards that discrimination; the median error, close to a factor of two whichever caprock-range subset is taken, is the more informative measure of what the estimate is worth on a rock already regarded as a plausible seal.

Nearly two thirds of the merged compilation was excluded after verification. 267 of 413 records were removed, each with a stated ground. The largest exclusion among the modeling-relevant data, a 25-record block with placeholder conditions and target pressures traceable to three distinct primary studies would have contributed 11 complete-feature records assigned to a single source lab. While sensitivity testing shows that incorporating this block leaves the transferable coefficient virtually unchanged (R^2 =+0.46), doing so fundamentally degrades the integrity of the source-aware validation by treating data from three independent laboratories as a single source fold.

The feature set omits the properties the physics identifies as first-order. No fluid-rock interaction property is available at the sample level, because interfacial tension and contact angle appear only as study-level constants. No pore-geometric descriptor is available at all, which is the most likely reason snap-off resists prediction. No quantitative mineralogy is present. Water saturation, which the literature shows raises breakthrough pressure exponentially, is not recorded. Sample length, which one study shows scales linearly with breakthrough pressure, is likewise unrecorded and therefore uncontrolled.

Confining pressure is excluded on the evidence of Section 5.7. The physical effect is real; this dataset cannot resolve it, because it records one stress value per study instead of a stress series per sample.

Hydrogen records is present in the dataset from two studies, one was excluded because it does not report the petrophysics its measurements would need. The hydrogen line of Chart BP-2 rests on three remaining records and on a measured hydrogen–brine interfacial tension of 71.6 mN/m from the same study; a second measurement in the literature, 51 to 64 mN/m, would lower the conversion factor from 0.19 to between 0.14 and 0.17, and the two are not reconciled by any record in the compilation. That is materially better than the interfacial-tension ratio alone, and it is not a validated hydrogen framework.

No prediction the framework returns carries its own uncertainty interval. Every band quoted is a population band, spread across seven orders of magnitude of permeability and applied uniformly, so a rock at the edge of the feature range is given the same factor as one in the middle although it is the less well constrained.

Finally, the mercury conversion is checked against five same-plug comparisons in the compilation, of which two fail, by a factor of ten in one direction and almost four in the other, and it carries two assumptions, contact angle and saturation convention, each capable of moving the estimate by a factor of two.

6.11 Recommendations for further work

Additional independent sources are the most likely route to a better estimate, since the learning curve was still rising when the data ran out, even if its last steps were small. What is needed is not more records but more laboratories and more rock types, and adding studies is more likely to help than adding records within the existing eight.

Per-sample measurement of interfacial tension and contact angle would let the fluid–rock interaction enter the feature space instead of being handled as a conversion factor after the fact. This is the single change that would most alter the modelling, and it is inexpensive relative to a breakthrough test.

A pore-geometric descriptor. Snap-off prediction needs a mercury-injection pore-throat distribution reported alongside the flow-through measurement, the closest routine measurement to the geometry it depends on. Without it, snap-off can only be bounded.

Core-scale hydrogen breakthrough measurements with accompanying petrophysics. Measurements of this kind have begun to appear and are the strongest immediate extension of this dataset. What most of them lack, and what the next study should supply, is the porosity and permeability of each plug alongside the threshold pressure, measured on brine-saturated cores, together with the hydrogen–brine interfacial tension measured at the test conditions.

Publication practice. Five reporting practices would raise the value of future compilations more cheaply than any new measurement: the pressure step of a step-by-step test, the confining stress for every sample, the saturation convention behind any mercury reading, the sample length, and the sample-level provenance of every record. The verification undertaken here found errors in ten of fourteen retained source blocks and excluded ten blocks of records; most of those errors would have been impossible had the original compilation carried a source identifier per record.

Publish the verified dataset. The field lacks a benchmark: each new method is applied to a different compilation with author-defined screening, which makes relative performance impossible to judge. The dataset delivered with this thesis carries a DOI and a verification status for every verified record and a plug identifier for every specimen. It is offered as a benchmark on which future methods can be compared on equal terms.

6.12 Conclusions

This investigation evaluated whether caprock sealing pressure can be reliably predicted from routinely measured petrophysical properties, assessing the operational value of such estimates. Ultimately, the finding is best defined as a boundary of predictability, delineating which capillary sealing quantity can be predicted, from which laboratory inputs, and to what degree of transferable accuracy across different predictive model.

Breakthrough pressure can be predicted with modest accuracy. A source-validated permeability-only correlation, refitted to the data, estimates it with a coefficient of determination of 0.46 on studies it has never seen, within a median factor of 2.6 over the full range and of 1.9 on the rocks that would be assessed as caprocks; adding porosity does not raise that coefficient. The correlation's record-level interval excludes zero, though a source-level bootstrap leaves every interval spanning zero. On the ultra-tight class, the median error is a factor of 1.7.

No model transfers better than a refitted permeability-only correlation. Both tree ensembles fit their training data better and transfer worse; a two-feature linear regression matches the correlation's transferable coefficient without exceeding it, and above 20 bar the correlation has the smaller median error. The framework therefore adopts the refitted permeability-only law as its predictive equation and keeps porosity only as a refinement in the estimation chart, where it lowers the full-range median error from a factor of 2.6 to 2.0. Selecting a source-validated, single-parameter permeability correlation over a complex ensemble preserves analytical simplicity, maximizing its operational value for field assessments.

Ultimately, the primary contribution of this work lies in establishing methodological guidelines, source-aware validation framework and a verified dataset benchmark for petrophysical dataset handling.

That accuracy is well below what the published literature reports, and the difference appears to stem mainly from validation protocols. Random-split validation allows a model to recognize studies it has partly seen. This is the principal methodological result of the work, and it is likely to apply to other compiled petrophysical datasets.

The overall findings are unaffected by individual screening criteria, as reversing all seven exclusions alters the cross-validation score only within a narrow band of 0.42 to 0.52. Performance is instead governed by dataset range; excluding five ambient-condition samples drops the R^2 to +0.18, while withholding one of the two largest source studies reduces it to +0.27. Data auditing therefore establishes strict provenance and physical interpretability, as source-aware validation depends on authentic source assignment and physically consistent inputs.

Data provenance constitutes a primary finding of this. Systematic verification against primary publications revealed that nearly two-thirds of the initial compiled records failed to meet data-quality standards. Most notably, excluding a 25-record unviable block removed what would have otherwise entered the modeling as the largest single training source. Whereas an unverified iteration of this dataset produced out-of-fold performance that collapsed below zero upon withholding its dominant source, the verified framework remains resilient; withholding either of the two largest primary sources reduces model performance without causing collapse

Snap-off pressure could not be predicted from these inputs by any method tested. The interpretation is physical; the pore-body-to-throat aspect ratio that governs it, is not encoded in bulk transport measurements.

The practical output of this thesis is therefore not a singular predictive model, but a tiered estimation framework that matches the analytical route to available petrophysical measurements—with each tier offering an explicit statement of its achievable accuracy, backed by a verified, traceable benchmark dataset to guide future research.

References

Ali, M., Yekeen, N., Pal, N., Keshavarz, A., Iglauer, S. & Hoteit, H. (2021). Influence of pressure, temperature and organic surface concentration on hydrogen wettability of caprock. Energy Reports, 7, 5988–5996. https://doi.org/10.1016/j.egyr.2021.09.016

Al-Mukainah, H., Al-Yaseri, A., Yekeen, N., Al Hamad, J. & Mahmoud, M. (2022). Wettability of shale–brine–H_2 system and H_2-brine interfacial tension for assessment of the sealing capacities of shale formations during underground hydrogen storage. Energy Reports, 8, 8830–8843. https://doi.org/10.1016/j.egyr.2022.07.004

Amaefule, J.O., Altunbay, M., Tiab, D., Kersey, D.G. & Keelan, D.K. (1993). Enhanced reservoir description: using core and log data to identify hydraulic (flow) units and predict permeability in uncored intervals/wells. SPE Annual Technical Conference, SPE-26436-MS.

Amann, A., Waschbüsch, M., Bertier, P., Busch, A., Krooss, B.M. & Littke, R. (2011). Sealing rock characteristics under the influence of CO_2. Energy Procedia, 4, 5170–5177. https://doi.org/10.1016/j.egypro.2011.02.494

Bachu, S. & Bennion, D.B. (2009). Interfacial tension between CO_2, freshwater and brine in the range of pressure from (2 to 27) MPa, temperature from (20 to 125) °C, and water salinity from (0 to 334 000) mg·L^{-1}. Journal of Chemical & Engineering Data, 54(3), 765–775. https://doi.org/10.1021/je800529x

Benlalam, N., Vasile, N.S., Peter, C., Rocca, V., Viberti, D. & Pirri, F. (2026). Experimental investigation of hydrogen breakthrough pressure for a clay-rich caprock. International Journal of Hydrogen Energy, 241, 155462. https://doi.org/10.1016/j.ijhydene.2026.155462

Bergen, K.J., Johnson, P.A., de Hoop, M.V. & Beroza, G.C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433), eaau0323. https://doi.org/10.1126/science.aau0323

Boulin, P.F., Bretonnier, P., Vassil, V., Samouillet, A., Fleury, M. & Lombard, J.M. (2013). Sealing efficiency of caprocks: experimental investigation of entry pressure measurement methods. Marine and Petroleum Geology, 48, 20–30. https://doi.org/10.1016/j.marpetgeo.2013.07.010

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, B., Li, X., Zhang, W., Wang, Y. & Liu, S. (2024). Breakthrough pressure of caprock: a review. Earth-Science Reviews, 252, 104732. https://doi.org/10.1016/j.earscirev.2024.104732

Chiquet, P., Broseta, D. & Thibeau, S. (2007). Wettability alteration of caprock minerals by carbon dioxide. Geofluids, 7(2), 112–122. https://doi.org/10.1111/j.1468-8123.2007.00168.x

Egermann, P., Lombard, J.M. & Bretonnier, P. (2006). A fast and accurate method to measure threshold capillary pressure of caprocks under representative conditions. International Symposium of the Society of Core Analysts, SCA2006-07, Trondheim.

Espinoza, D.N. & Santamarina, J.C. (2017). CO_2 breakthrough — caprock sealing efficiency and integrity for carbon geological storage. International Journal of Greenhouse Gas Control, 66, 218–229. https://doi.org/10.1016/j.ijggc.2017.09.019

Fleury, M., Pironon, J., Le Nindre, Y.M., Bildstein, O., Berne, P., Lagneau, V., Broseta, D., Pichery, T., Fillacier, S., Lescanne, M. & Vidal, O. (2011). Evaluating sealing efficiency of caprocks for CO_2 storage: an overview of the Geocarbone Integrity Program. Energy Procedia, 4, 5227–5234. https://doi.org/10.1016/j.egypro.2011.02.501

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, S., Wei, N., Li, X., Wang, Y. & Wang, Q. (2014). Cap rock CO_2 breakthrough pressure measurement apparatus and application in Shenhua CCS project. Energy Procedia, 63, 4766–4772. https://doi.org/10.1016/j.egypro.2014.11.507

Hao, S., Cao, J., Zhang, H., Liu, Y., Liang, H. & Li, M. (2021). Breakthrough pressure prediction based on neural network model. Geofluids, 2021, 6154468. https://doi.org/10.1155/2021/6154468

Hildenbrand, A., Schlömer, S. & Krooss, B.M. (2002). Gas breakthrough experiments on fine-grained sedimentary rocks. Geofluids, 2(1), 3–23. https://doi.org/10.1046/j.1468-8123.2002.00031.x

Hildenbrand, A., Schlömer, S., Krooss, B.M. & Littke, R. (2004). Gas breakthrough experiments on pelitic rocks: comparative study with N_2, CO_2 and CH_4. Geofluids, 4(1), 61–80. https://doi.org/10.1111/j.1468-8123.2004.00073.x

Hubert, L. & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218. https://doi.org/10.1007/BF01908075

Ito, D., Akaku, K., Okabe, T., Takahashi, T. & Tsuji, T. (2011). Measurement of threshold capillary pressure for seal rocks using the step-by-step approach and the residual pressure approach. Energy Procedia, 4, 5211–5218. https://doi.org/10.1016/j.egypro.2011.02.499

Jia, J., Fan, W., Lu, Y., Feng, H. & Yang, B. (2026). Experimental investigation of the controlling factors of CO_2 breakthrough pressure and the pore–throat evolution. Energies, 19(7), 1767. https://doi.org/10.3390/en19071767

Karami, S., Ahmadi, M., Zendehboudi, S. & Chatzis, I. (2025). From empirical to intelligent: ML-enhanced parachor modelling of gas–water interfacial tension. Results in Engineering, 28, 107554. https://doi.org/10.1016/j.rineng.2025.107554

Kawaura, K., Akaku, K., Nakano, M., Ito, D., Takahashi, T. & Kiriakehata, S. (2013). Examination of methods to measure capillary threshold pressures of pelitic rock samples. Energy Procedia, 37, 5411–5418. https://doi.org/10.1016/j.egypro.2013.06.459

Kim, H., Vilarrasa, V. & Makhnenko, R.Y. (2025). Laboratory-scale assessment of CO_2 sealing potential of heterogeneous caprock. Water Resources Research, 61. https://doi.org/10.1029/2025WR040339

Kim, S. & Santamarina, J.C. (2013). CO_2 breakthrough and leak-sealing — experiments on shale and cement. International Journal of Greenhouse Gas Control, 19, 471–477. https://doi.org/10.1016/j.ijggc.2013.10.011

Klimkowski, Ł. & Smulski, R. (2012). Laboratory method to measure sealing capacity of caprocks. Archives of Mining Sciences, 57(2), 471–481. https://doi.org/10.2478/v10267-012-0031-4

Kolodzie, S. (1980). Analysis of pore throat size and use of the Waxman–Smits equation to determine OOIP in Spindle Field, Colorado. SPE Annual Technical Conference, SPE-9382-MS.

Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.

Lenormand, R., Zarcone, C. & Sarr, A. (1983). Mechanisms of the displacement of one fluid by another in a network of capillary ducts. Journal of Fluid Mechanics, 135, 337–353. https://doi.org/10.1017/S0022112083003110

Li, Q., Gates, I.D. & Hejazi, S.H. (2025). Experimental investigation of caprock sealing capacity for underground hydrogen storage. International Journal of Hydrogen Energy, 165, 150858. https://doi.org/10.1016/j.ijhydene.2025.150858

Li, S., Dong, M., Li, Z., Huang, S., Qing, H. & Nickel, E. (2005). Gas breakthrough pressure for hydrocarbon reservoir seal rocks: implications for the security of long-term CO_2 storage in the Weyburn field. Geofluids, 5(4), 326–334. https://doi.org/10.1111/j.1468-8123.2005.00125.x

Liu, P., Song, R., Fu, B., Liu, J. & Yang, C. (2025). Study on breakthrough pressure of H_2, CH_4, and CO_2 in tight caprock for underground hydrogen storage. Energy Storage, 7(8), e70280. https://doi.org/10.1002/est2.70280

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.

Makhnenko, R.Y., Vilarrasa, V., Mylnikov, D. & Laloui, L. (2017). Hydromechanical aspects of CO_2 breakthrough into clay-rich caprock. Energy Procedia, 114, 3219–3228. https://doi.org/10.1016/j.egypro.2017.03.1453

Medina, C.R., Rupp, J.A., Lahann, R. & Eldridge, J. (2019). Evaluation of caprock integrity of the Upper Ordovician units within the CarbonSAFE prefeasibility study region, Subtask 4.3 — geological characterization. Indiana Geological & Water Survey, Report DOE-FE0029445-11. OSTI 1524067.

Pittman, E.D. (1992). Relationship of porosity and permeability to various parameters derived from mercury injection-capillary pressure curves for sandstone. AAPG Bulletin, 76(2), 191–198.

Rezaeyan, A., Tabatabaei-Nejad, S.A., Khodapanah, E. & Kamari, M. (2015). Parametric analysis of caprock integrity in relation with CO_2 geosequestration. Greenhouse Gases: Science and Technology, 5(6), 714–731. https://doi.org/10.1002/ghg.1516

Rocca, V. (2021). The sealing efficiency of cap rocks — laboratory tests and an empirical correlation. Geoingegneria Ambientale e Mineraria, LVIII (1), 41–48. https://doi.org/10.19199/2021.1.1121-9041.041

Roof, J.G. (1970). Snap-off of oil droplets in water-wet pores. SPE Journal, 10(1), 85–90. https://doi.org/10.2118/2504-PA

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Skerbisch, L., Misch, D., Drews, M., Arnberger, K., Schuller, V., Zamolyi, A., Hantschel, T., Palmowski, D. & Kleine, A. (2025). Caprock integrity: a critical factor for carbon capture and storage in the Vienna Basin. International Journal of Greenhouse Gas Control, 146, 104434. https://doi.org/10.1016/j.ijggc.2025.104434

Span, R. & Wagner, W. (1996). A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. Journal of Physical and Chemical Reference Data, 25(6), 1509–1596. https://doi.org/10.1063/1.555991

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15(1), 72–101.

Thomas, L.K., Katz, D.L. & Tek, M.R. (1968). Threshold pressure phenomena in porous media. SPE Journal, 8(2), 174–184. https://doi.org/10.2118/1816-PA

Tonnet, N., Mouronval, G., Chiquet, P. & Broseta, D. (2011). Petrophysical assessment of a carbonate-rich caprock for CO_2 geological storage purposes. Energy Procedia, 4, 5422–5429. https://doi.org/10.1016/j.egypro.2011.02.527

Tukey, J.W. (1977). Exploratory data analysis. Addison-Wesley.

Vespo, V.S., Messori, A. & Musso, G. (2024). Threshold capillary pressure of caprocks for CO_2 storage: numerical insight on the dynamic and residual method. International Journal of Greenhouse Gas Control, 135, 104135.

Wollenweber, J., Alles, S., Busch, A., Krooss, B.M., Stanjek, H. & Littke, R. (2010). Experimental investigation of the CO_2 sealing efficiency of caprocks. International Journal of Greenhouse Gas Control, 4(2), 231–241. https://doi.org/10.1016/j.ijggc.2010.01.003

Wu, H., Zhang, M., Wu, Z. & Liu, Z. (2026). Shale cap breakthrough pressure prediction method based on machine learning. Applied Sciences, 16, 4675. https://doi.org/10.3390/app16104675

Xu, C., Fu, L., Lin, T., Li, W. & Ma, S. (2022). Machine learning in petrophysics: advantages and limitations. Artificial Intelligence in Geosciences, 3, 157–161. https://doi.org/10.1016/j.aiig.2022.11.004

Zhang, C. & Yu, Q. (2016). The effect of water saturation on methane breakthrough pressure: an experimental study on the Carboniferous shales from the eastern Qaidam Basin, China. Journal of Hydrology, 543, 512–520. https://doi.org/10.1016/j.jhydrol.2016.11.003

Zhao, Y. & Yu, Q. (2018). Effect of CH_4 on the CO_2 breakthrough pressure and permeability of partially saturated low-permeability sandstone in the Ordos Basin, China. Journal of Hydrology, 556, 732–748. https://doi.org/10.1016/j.jhydrol.2017.11.030

Appendix A — Source publications and dataset provenance

A.1 Retained sources

Table A.1 lists the fourteen publications contributing to the 146-record verified foundation. Every verified record carries the DOI in its `Source_DOI` column and a verification status in `Provenance_Status`, so that any value can be traced to a published table. Records excluded for lack of a traceable source are retained in the delivered file with no DOI and with their exclusion ground recorded.

Table A.1 — Source publications of the verified dataset.

Code	Publication	Doi / identifier	Records	Plugs	Method	Fluids	Lithology
A.HILD	Hildenbrand, Schlömer, Krooss & Littke (2004), Geofluids 4, 61–80	10.1111/j.1468-8123.2004.00073.x	35	23	residual	CH_4, CO_2, N_2	clay, mudstone, shale
EVALOF	Medina, Rupp, Lahann & Eldridge (2019), Indiana Geological & Water Survey, CarbonSAFE Subtask 4.3	OSTI 1524067	28	28	mercury injection	Hg	limestone, shale
DAISITO	Ito, Akaku, Okabe, Takahashi & Tsuji (2011), Energy Procedia 4, 5211–5218	10.1016/j.egypro.2011.02.499	12	3	step-by-step, residual, mercury	Hg, N_2	mudstone
AFAST	Egermann, Lombard & Bretonnier (2006), SCA2006-07, Trondheim	SCA2006-07	10	5	step-by-step, residual	N_2	carbonate, chalk, sandstone, shale
GASBREA	Li, Dong, Li, Huang, Qing & Nickel (2005), Geofluids 5, 326–334	10.1111/j.1468-8123.2005.00125.x	9	7	step-by-step	CO_2, N_2	anhydrite, mudstone
LIET2025	Li, Gates & Hejazi (2025), Int. J. Hydrogen Energy 165, 150858	10.1016/j.ijhydene.2025.150858	9	3	step-by-step	CH_4, H_2, N_2	shale
KAWA	Kawaura, Akaku, Nakano, Ito, Takahashi & Kiriakehata (2013), Energy Procedia 37, 5411–5418	10.1016/j.egypro.2013.06.459	9	1	step-by-step, residual, mercury	CO_2, Hg, N_2	mudrock
REZA2015	Rezaeyan, Tabatabaei-Nejad, Khodapanah & Kamari (2015), Greenh. Gases Sci. Technol. 5, 714–731	10.1002/ghg.1516	9	6	step-by-step, residual	CH_4, CO_2, N_2	anhydrite, shale
GASBSEDI	Hildenbrand, Schlömer & Krooss (2002), Geofluids 2, 3–23	10.1046/j.1468-8123.2002.00031.x	6	6	residual	N_2	clay, mudstone
MAKH2017	Makhnenko, Vilarrasa, Mylnikov & Laloui (2017), Energy Procedia 114, 3219–3228	10.1016/j.egypro.2017.03.1453	6	6	residual	CO_2	Opalinus Clay
WOLLET	Wollenweber, Alles, Busch, Krooss, Stanjek & Littke (2010), IJGGC 4, 231–241	10.1016/j.ijggc.2010.01.003	5	2	residual	CO_2	limestone, marlstone
SHUAGAO	Gao, Wei, Li, Wang & Wang (2014), Energy Procedia 63, 4766–4772	10.1016/j.egypro.2014.11.507	4	4	residual	CO_2	mudstone

Code	Publication	Doi / identifier	Records	Plugs	Method	Fluids	Lithology
METOWYKO	Klimkowski & Smulski (2012), Archives of Mining Sciences 57(2), 471–481	10.2478/v10267-012-0031-4	3	3	step-by-step	CO_2	shale
ALEXAMAN	Amann, Waschbüsch, Bertier, Busch, Krooss & Littke (2011), Energy Procedia 4, 5170–5177	10.1016/j.egypro.2011.02.494	1	1	step-by-step, residual	CO_2	marl

Three entries need a note. KAWA's nine records are measurements on a single mudrock plug. A second block of six records had been held in the compilation under a separate code (KAWA2) and attributed to the same publication on the resemblance of its petrophysics; on re-reading, Kawaura et al. (2013) reports that one plug only and none of the six values, so the block is now excluded as unattributable (Table A.3). LIET2025's nine records are three plugs (E, W and M), each tested with nitrogen, methane and hydrogen (Li et al., 2025, Table 2). WOLLET's five records are repeat carbon-dioxide exposures on two plugs, not five samples, and the source reports sealing capacity falling by a third to a half over successive exposures, so the five values are not replicates of a fixed quantity.

A.2 Corrections applied during verification

Ten of these corrections were made to records in the compilation as received (§3.2); the other two — LIET2025 and SHUAGAO — were made to records added during the enrichment of §3.3

Table A.2 — Value corrections, each made against the published table.

Source	Correction
GASBREA	reference title named the wrong publication; the values match Li et al. (2005) Table 2 exactly. Label corrected, data unchanged
METOWYKO	three breakthrough pressures lay outside the ranges the paper reports; set to the range midpoints. Confining pressure corrected from 150 to 140 bar
ALEXAMAN	a breakthrough value was held in the snap-off column, and a second value appeared nowhere in the source; both corrected to the published figures. The CS_12 snap-off value of 6.3 bar had been entered twice; the second entry is marked as a duplicate
A.HILD	one porosity corrected from 0.040 to 0.044; one plug carried two different porosity measures from the same table, and the mercury-injection value was adopted throughout for consistency
EVALOF	permeability was the mercury-derived Katz estimate, which the source itself reports as up to four orders of magnitude above measurement; replaced with the crushed-core values from the same report for nineteen samples and removed where no measurement exists. Entry pressure changed from the first-intrusion reading to the twenty-per-cent saturation reading, for the reason given in Section 3.5
DAISITO	one converted mercury value present in the source was missing from the compilation and has been added. The Core 3 step-by-step value, also missing, has been added at 28.75 bar, the midpoint of the 2750–3000 kPa range in Ito et al. (2011) Table 2, the convention used for Cores 1 and 2. Core 1 was saturated with pure water and not brine (Ito et al., 2011, §3.2), and its wetting phase is recorded as water

Source	Correction
WOLLET	*porosity corrected by a factor of 100, from a per-cent value recorded as a fraction; plug identity restored, so that five records are correctly attributed to two plugs*
AFAST	*plug identity rebuilt: the ten records are a step-by-step and a residual value on each of five samples, R1 to R5 (Egermann et al., 2006, Table 2)*
KAWA	*plug identity rebuilt: all records are measurements on one mudrock plug (Kawaura et al., 2013). The six records formerly held under KAWA2 were re-examined against Kawaura et al. (2013) and Ito et al. (2011); neither reports them, and they are excluded as unattributable*
REZA2015	*the three residual records re-keyed to the plugs of the step-by-step measurements they were made on (Rezaeyan et al., 2015, Table 3)*
LIET2025	*plug identity rebuilt: the nine records are three plugs, E, W and M, each tested with the three gases (Li et al., 2025, Table 2), not nine plugs. The breakthrough-pressure column of the paper is used, not its threshold-pressure column. Confining pressure flagged: it was held 5 MPa above the injection pressure (Li et al., 2025, §2.2), so at breakthrough it equals the breakthrough pressure plus 50 bar, and it is never used as a model input*
SHUAGAO	*measurement type reassigned: Gao et al. (2014) name the quantity a carbon-dioxide breakthrough pressure but measure it by the pulse-decay residual method, so its four records are assigned to the snap-off (S2) target on the basis of the method, not the name (Section 3.2). Values unchanged*

One proposed correction was withdrawn on re-checking. A value that appeared to be a transcription slip proved to be a second measurement on a second fragment of the same plug, and both figures are published.

A.3 Blocks removed and the grounds for removal

Table A.3 — Blocks excluded from the verified foundation.

Block	Records	Ground
field-scale review compilation	*79*	*physics. Basin-scale sealing estimates over a formation-scale support volume, a different quantity from a plug measurement, with no petrophysics*
literature compilation added during enrichment (Hao et al.)	*55*	*traceability. Compiles sixteen publications (thirteen Table 1 rows), most already primary sources here, with values converted to a common basis*
mixed-source block with placeholder conditions	*25*	*data quality. Targets traceable to three studies; permeability off by a factor of 1000, temperature and confining pressure identical across the block at the critical point of carbon dioxide*
unattributable records	*16*	*traceability. Complete petrophysics but no identifiable source, so no grouping key can be assigned for source-aware validation*

Block	Records	Ground
records attributed by resemblance only (formerly KAWA2)	*6*	*traceability. Attributed to Kawaura et al. (2013) on the resemblance of their petrophysics; the publication reports one plug and none of these values, so no source can be assigned*
figure-derived values (Kim, Vilarrasa & Makhnenko, 2025)	*6*	*traceability. Evenly spaced values and facies labels indicate readings taken from a chart; no per-sample table located*
records without measured petrophysics (Benlalam et al., 2026)	*8*	*data quality. Porosity and permeability assumed, the source reporting permeability only as upper bounds*
crack-dominated measurements	*3*	*physics. The source states that the measured pressures imply pore radii far larger than the rock allows, and reports hairline cracks in every plug*
reservoir sandstone	*2*	*physics. The authors describe the samples as reservoir rock, outside the caprock scope*
disproved attribution	*2*	*traceability. The cited publication is a numerical study containing no laboratory measurement*

A.4 Sources identified but not retained

Liu et al. (2025), Energy Storage 7, e70280 reports hydrogen, methane and carbon-dioxide threshold pressures on tight caprock with full petrophysics. The cores were saturated with anhydrous kerosene and not brine. A gas–kerosene threshold is a different quantity from a gas–brine one, so the records are excluded on the wetting-phase criterion recorded in the Wetting_Phase column.

Jia et al. (2026), Energies 19, 1767 tested thirty-six Ordos caprock samples but publishes permeability per sample only, with breakthrough pressure given as regression curves. No per-sample pairing of features and target is recoverable.

Benlalam et al. (2026), Int. J. Hydrogen Energy 241, 155462 is retained in the dataset and cited for its hydrogen result, but excluded from modelling because it reports permeability only as upper bounds and no porosity.

A.5 The delivered dataset

AllData_verified_corrected_v2.csv holds 413 records, 348 of them non-duplicate, representing 298 plugs across 24 columns. It is produced from AllData_verified_corrected.csv by the script apply_fixes_v2.py delivered alongside it, which makes two changes and records each in the data_flag column: the six KAWA2 records are marked unattributable (ML_exclude, Provenance_Status, Source_DOI and Source_Citation updated), and the nine LIET2025 plug identifiers are rebuilt from nine to three. Beyond the measured quantities the file carries six attributes that make the verification auditable:

Column	*Purpose*
Plug_Id	*identifies the physical specimen, so repeat measurements are distinguishable from separate samples*
Source_DOI	*DOI or persistent identifier for every verified record*
Provenance_Status	*how far the record was checked against its primary source*
ML_exclude	*the exclusion ground, so that every screening decision is visible in the data*
P_semantics	*whether the pressure column holds applied confining stress or in-situ formation pressure*
Wetting_Phase	*brine or water against a dry plug, so the mercury and kerosene distinctions are enforceable*

All statistics, figures and models in Chapters 3 to 6 are reproducible from this file with the scripts delivered alongside it, run in the order `pipeline.py` (screening funnel and the S1, S2 and mercury populations), `analysis.py` (Tables 5.2 and 5.5, the fitted equations, bootstrap intervals, in-sample importance and learning curve), `ch346.py` (Tables 3.3, 3.5, 3.7, 3.9, 6.1, 6.2 and 6.4, the conversion factors and the source-removal test), `extras.py` (Tables 3.8 and 5.3, same-plug pairs, out-of-fold importance, completeness), `supplementary.py` (Tables 3.4, 3.6, 4.1, 4.3, 5.4, 5.6 and 6.3), `seed_sensitivity.py` (the seed check of Section 5.3) and `new_analyses.py` (the published-Thomas and Winland baselines of Section 5.5, the leave-one-class-out result of Section 6.4, the full source-removal test of Table 6.1a, and the random-split importance of Section 5.6), at a fixed random seed of 42. No pressure window is applied in `pipeline.py`; the 50–220 bar window appears only as an ablation row in Table 5.4.

Appendix B — Supporting figures

These figures support statements made in Chapter 3 and are placed here so that the chapter carries only the figures its argument depends on. Each is referenced from the text at the point it bears on.

B.1 The compilation and its completeness

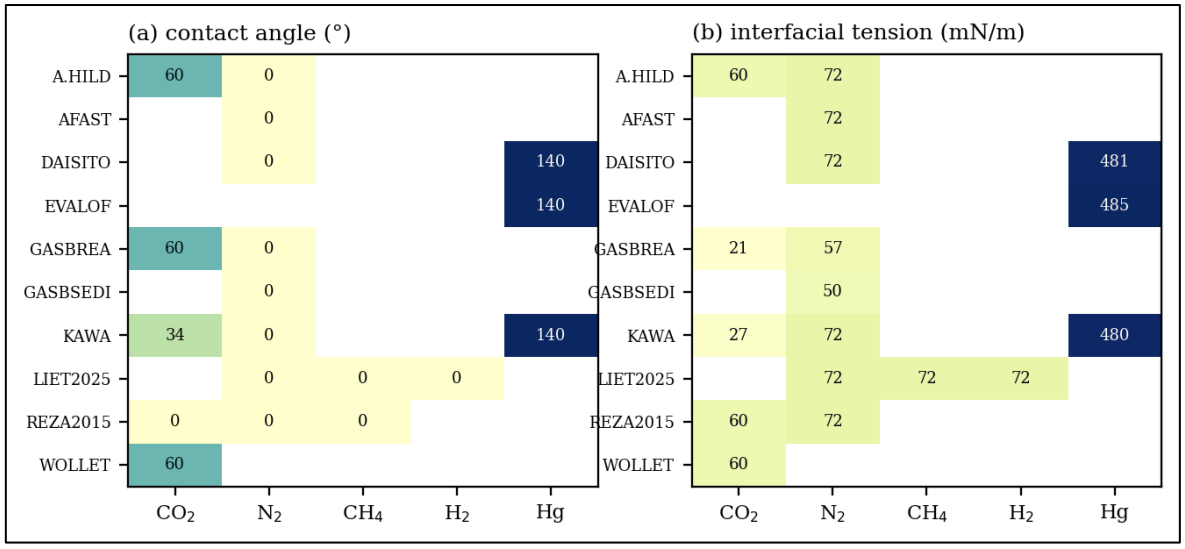


Figure B.1 — Contact angle and interfacial tension carried by each publication, by fluid.

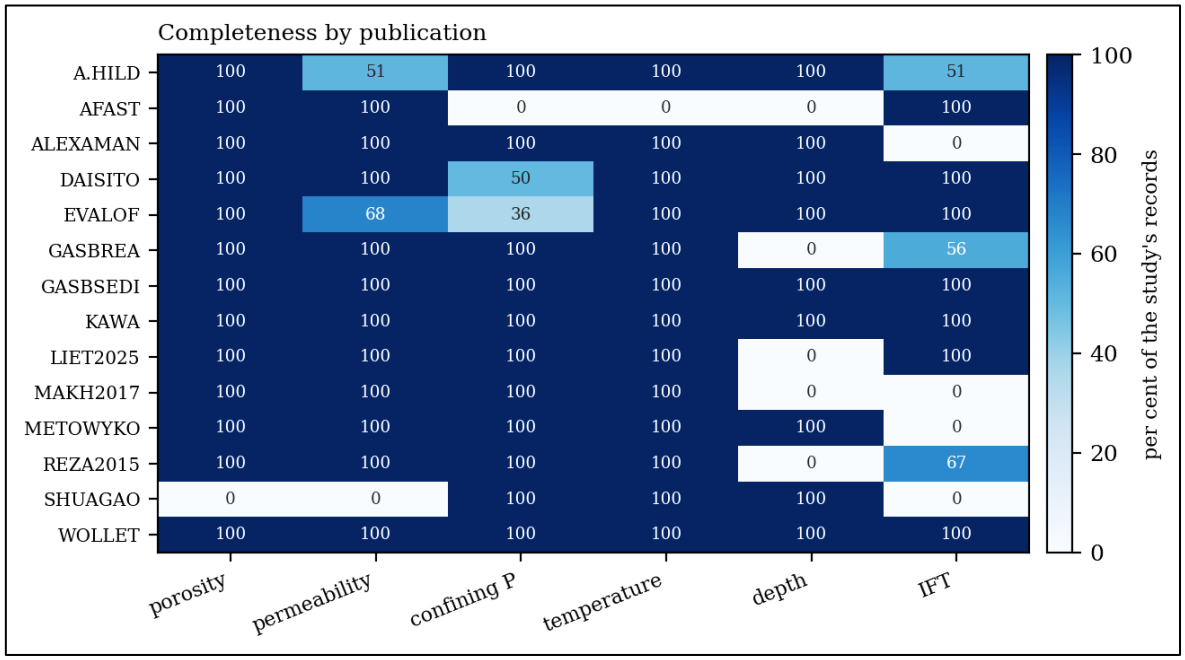


Figure B.2 — Completeness by publication, as the percentage of each study's records reporting each quantity. The gaps form horizontal bands, one study at a time.

Missingness follows the study and not the sample. Porosity, temperature and depth are reported by each publication either for all of its samples or for none, and permeability and confining pressure in 86 per cent of publications.

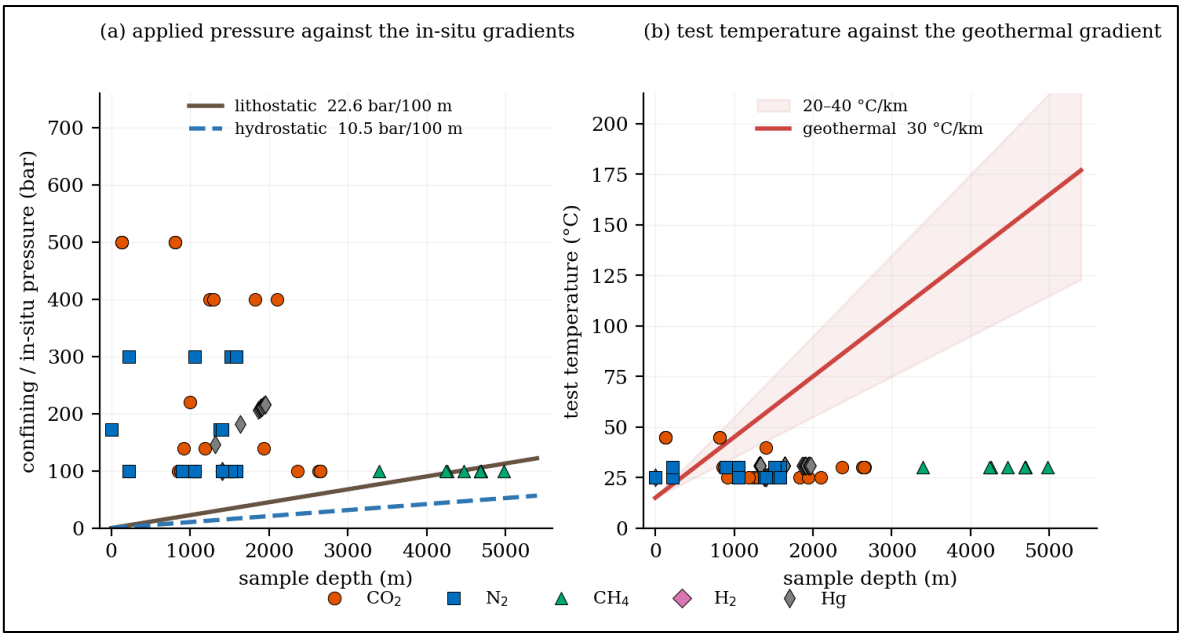


Figure B.3 — Confining pressure and test temperature against sample depth, with the in-situ gradients.

Neither confining pressure nor temperature has a usable relationship with sample depth. The figure makes the point against the gradients that would be used to impute them. In panel (a) the applied confining pressures scatter across the lithostatic and hydrostatic lines without following either: several studies apply 100 bar to samples from 200 m and from 2600 m alike, while one applies 500 bar to a sample from 800 m. In panel (b) the test temperatures sit almost entirely below a 30 °C per kilometre geothermal gradient, and most cluster at 25 or 30 °C irrespective of depth, because they are laboratory conditions of convenience and not in-situ conditions.

This is why neither quantity can be imputed from depth. An imputation from a gradient would replace a laboratory setting, chosen by the experimenter, with an estimate of the in-situ state the sample was never tested at. It would also give every sample from one study the same substituted value, so a model could read that value as a marker of the study.

B.2 The targets across fluids and rock types

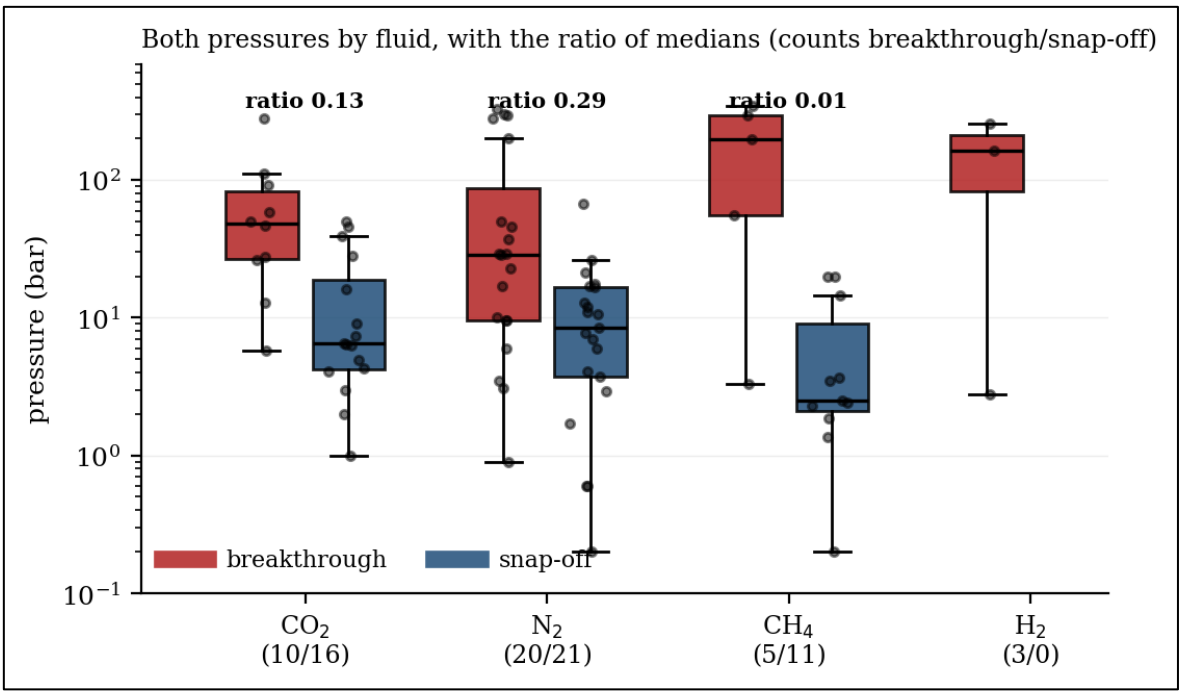


Figure B.4 — Both pressures by fluid, with the ratio of medians.

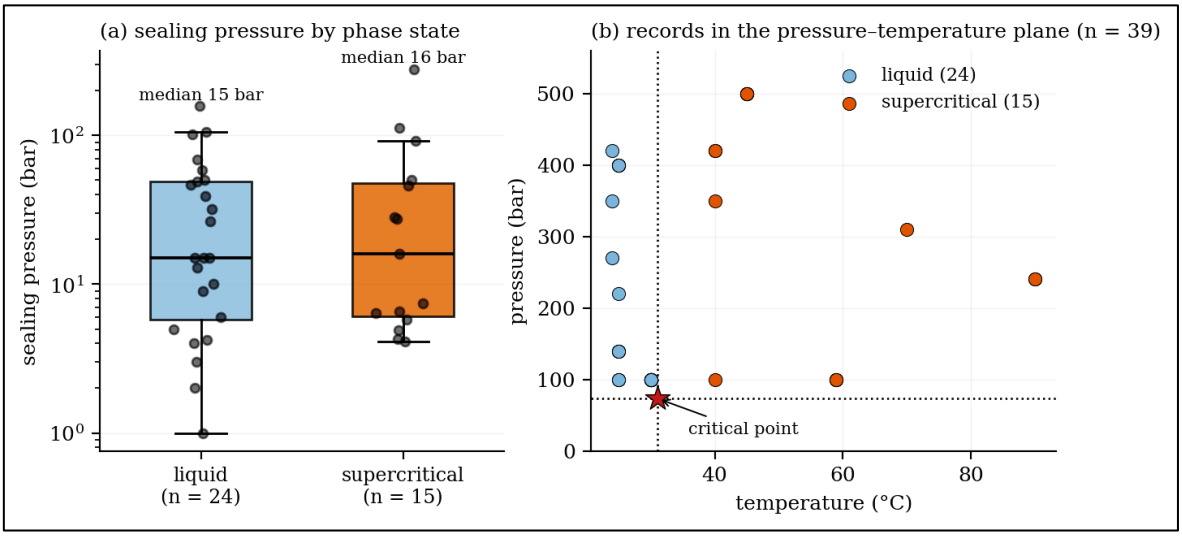


Figure B.5 — Sealing pressure by carbon-dioxide phase state, and the records in the pressure–temperature plane.

Twenty-four carbon dioxide records are liquid and fifteen supercritical. Panel (a) compares the sealing pressures of the two populations directly: the boxes overlap almost entirely and the medians differ by less than the spread within either group. Phase is therefore not carried as a feature.

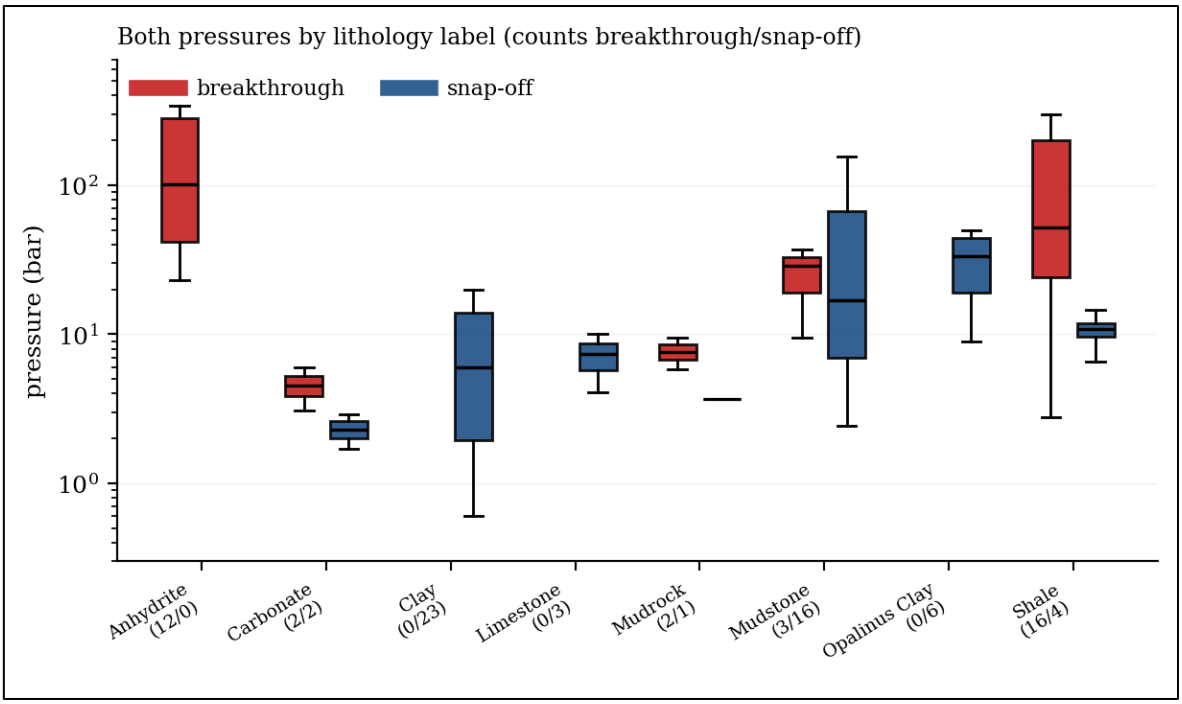


Figure B.6 — Breakthrough and snap-off pressure by lithology label, for labels with at least four records; counts in brackets.

Each lithology label pools different fluids and methods in different proportions, and the counts behind several labels are small.

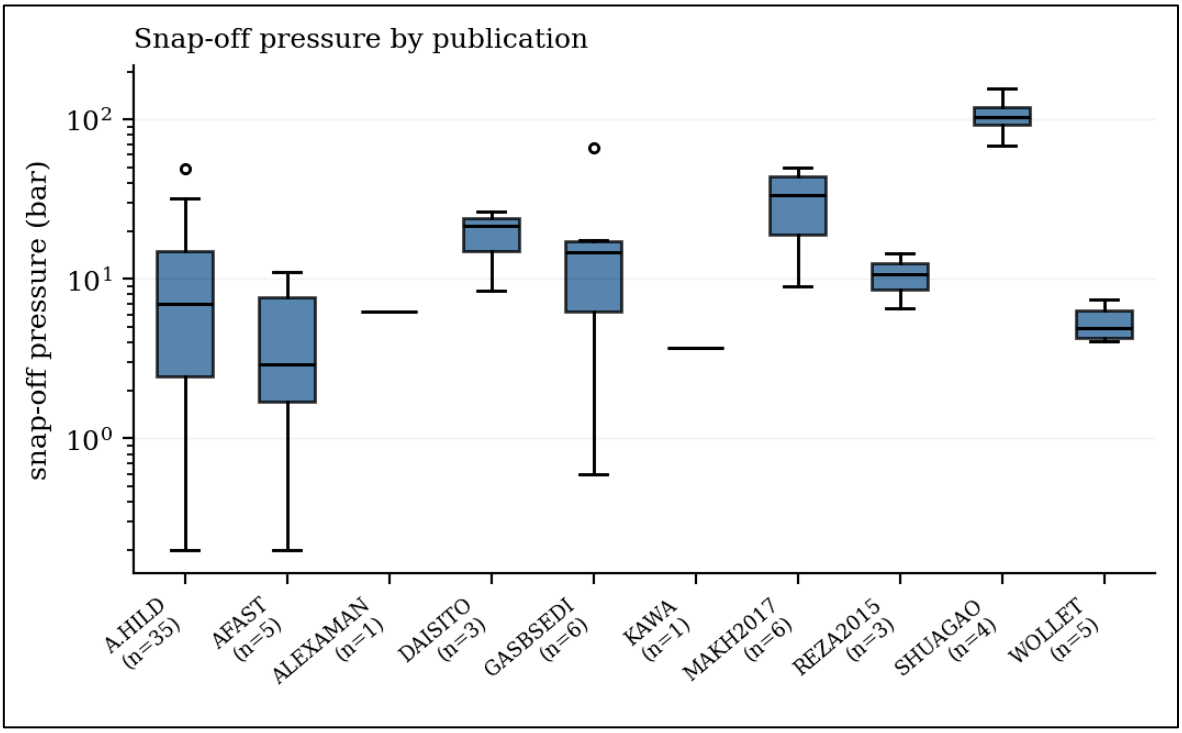


Figure B.7 — Snap-off pressure by publication, showing the range of each study and its median.

Each study spans a wide range of its own and the study medians overlap. Sixty-three per cent of the variance of snap-off pressure lies within studies.

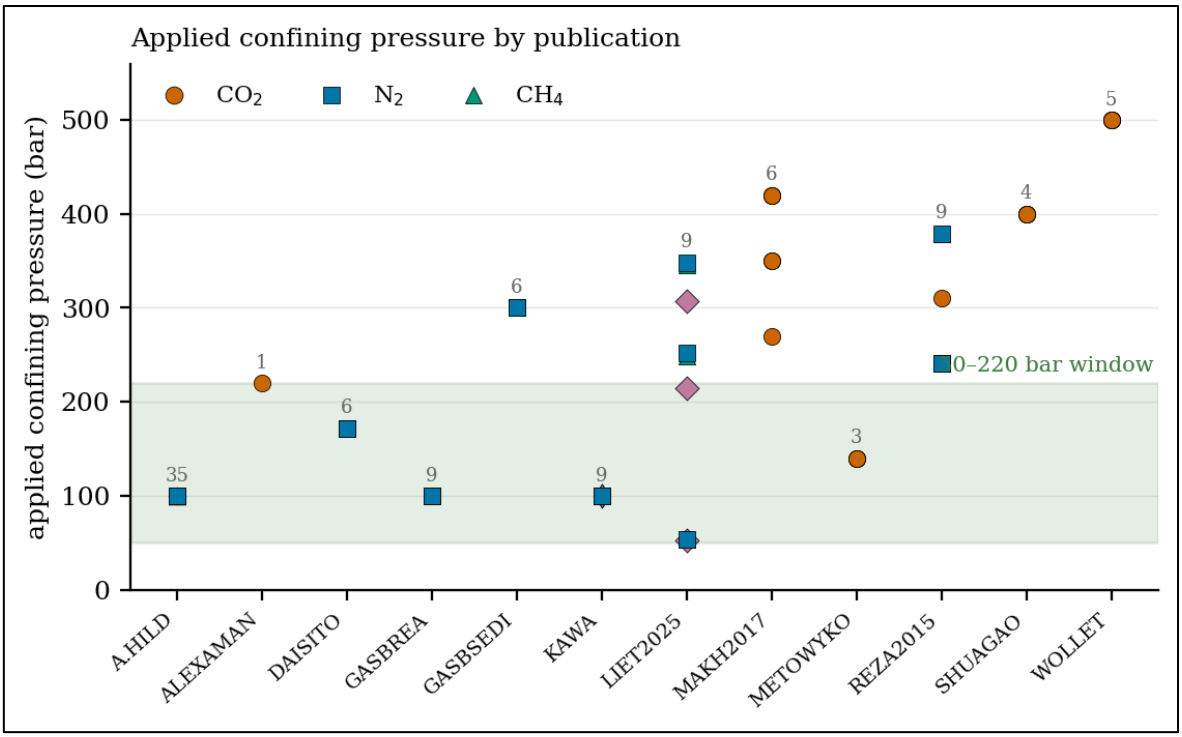


Figure B.8 — Applied confining pressure by publication, with the 50 to 220 bar window shaded.

Most publications apply a single confining pressure to all their samples. Two studies sit outside the 50 to 220 bar window entirely, one at ambient conditions and one between 241 and 379 bar.

B.3 The petrophysical grouping

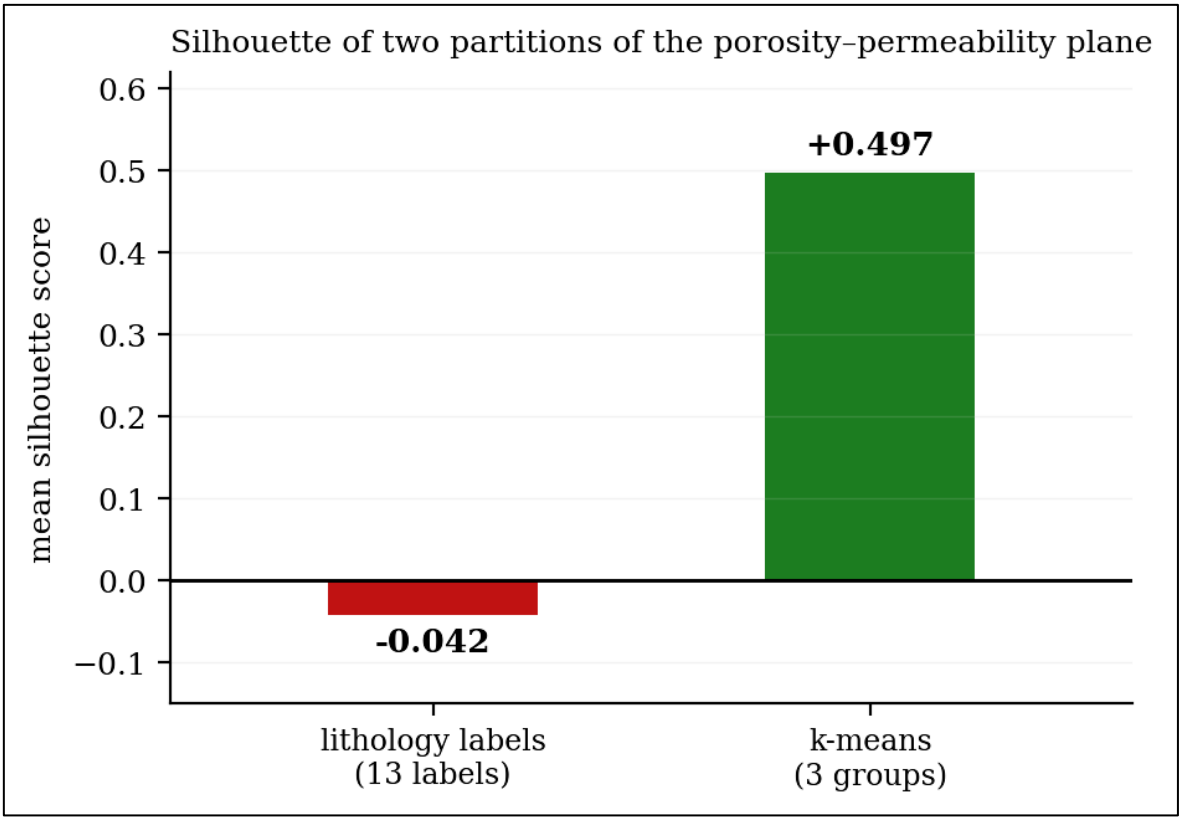


Figure B.9 — Silhouette of two partitions of the porosity–permeability plane.

The silhouette of the thirteen-label lithology partition in the porosity–permeability plane, with label case normalised, is −0.04, against +0.497 for the three k-means groups. A negative value means a sample is, on average, closer to rocks carrying other labels than to rocks carrying its own.

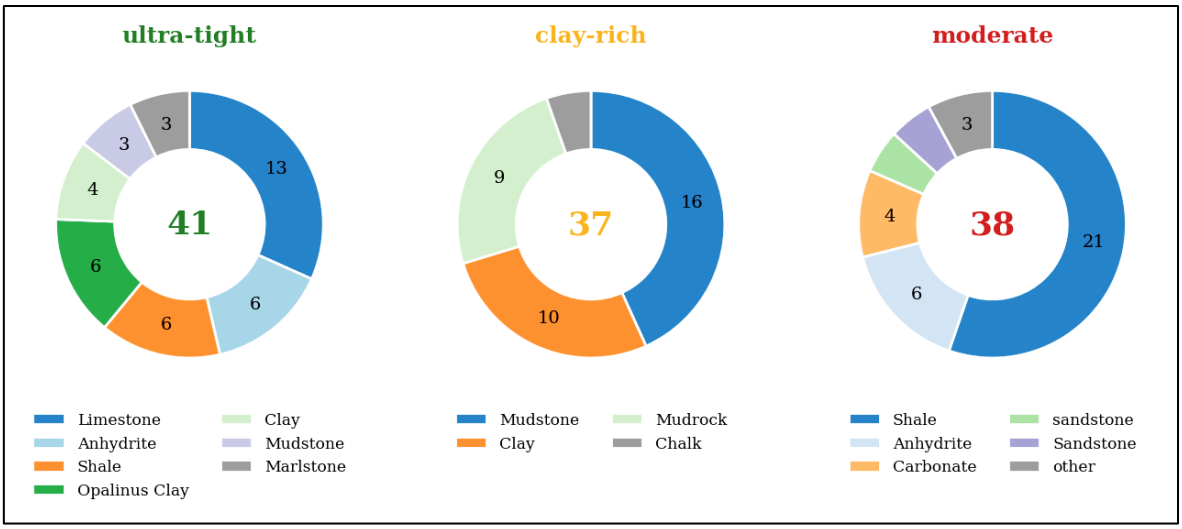


Figure B.10 — Lithological composition of each seal-quality group.

The groups are not lithological classes. The ultra-tight group is led by limestone and anhydrite, the clay-rich group by mudstone, clay and mudrock, and the moderate group by shale and carbonate. Each holds several labels, and the same label appears in more than one group.