IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Pharmaceutical Sciences

Vol 1, No 1, 2026

Physicochemical Determinants of Gram-Negative Penetration and the Limits of Machine Learning in Antibacterial Discovery: A Theoretical Analysis

IGWE Emmanuel Sunday[1], AMADI Daniel Chiemerie[2]

[1] Veristzon Global Research Lab; [2] Faculty of Pharmaceutical Sciences, University of Nigeria

ABSTRACT

Poor penetration and accumulation of compounds in Gram-negative bacteria is a principal reason that potent antibacterial leads fail, and machine-learning models are increasingly used to predict permeation and accumulation from chemical structure. This theoretical paper asks what such models can reasonably be expected to learn. The study is non-empirical and uses a conceptual research design. A corpus of peer-reviewed journal articles published between 2020 and 2026 was assembled across four literatures, namely outer-membrane permeation and porin-mediated uptake, efflux transport and its inhibition, compound accumulation measurement and the prediction tools built upon it, and machine-learning discovery of antibacterial compounds, with a supplementary search on resistance burden and diagnostic capacity in sub-Saharan Africa. Every source was verified against its Crossref publisher record for authorship, journal, volume, pagination and year, and screened for retractions, corrections and expressions of concern in both that record and the corresponding PubMed record; sources failing verification were excluded rather than corrected by inference. Analysis proceeded by classifying each measurement in routine use according to the processes it collapses, and by theory construction from the resulting pattern. Three findings are reported. First, every label obtainable at the scale machine learning requires is a composite in which influx across the outer membrane, active efflux, partitioning between cellular compartments and target engagement are summed into one number, so the separate contributions are not identifiable and mechanistic interpretations of such models are unlicensed. Second, because the relative weight of these processes varies with organism, growth state and assay, transfer failure is governed by shift in that mixture rather than by chemical distance, which explains why accumulation rules established in one species need not hold in another whose dominant route of entry differs. Third, the inference a model licenses is fixed by its held-out design rather than by its reported accuracy. The paper develops a four-layer framework relating the quantity of interest to the available label, the confounds between them and the validation design required, distinguishes chemotype, species and condition failure as modes with distinct remedies, and derives six falsifiable propositions and eleven reporting standards. It is recommended that performance be reported separately by split design, that condition-annotated and organism-resolved measurement be prioritised over further accumulation of heterogeneous data, and that model applicability for African health systems be established in locally circulating isolates rather than assumed from reference strains.

Keywords: antibacterial drug discovery; Gram-negative bacteria; permeability; compound accumulation; efflux

Physicochemical Determinants of Gram-Negative Penetration and the Limits of Machine Learning in Antibacterial Discovery: A Theoretical Analysis

1. INTRODUCTION

The organisms that carry the heaviest share of the resistance burden are predominantly Gram-negative. Deaths associated with bacterial antimicrobial resistance reached an estimated 4.71 million in 2021, and of all antibiotic classes it is resistance to carbapenems among Gram-negative organisms that has risen fastest; the same analysis estimates that a drug pipeline directed at Gram-negative pathogens could cumulatively avert some 11.1 million resistance-attributable deaths between 2025 and 2050 (Naghavi et al., 2024). In sub-Saharan Africa the projected toll is severe and surveillance to characterise it is thin (Totaro et al., 2025).

The scientific obstacle is long established and specific. Gram-negative bacteria are enclosed by an outer membrane whose asymmetric glycolipid leaflet excludes large and hydrophobic molecules, and whose hydrophilic channels admit only compounds of restricted size, charge and shape; what does cross is then subject to active expulsion by broad-specificity efflux transporters. The two mechanisms act together rather than additively, so that small changes in either can produce large changes in how much compound reaches an intracellular target (Zgurskaya, Walker, et al., 2021). Compounds potent against isolated targets therefore routinely fail against whole Gram-negative cells, and the failure is one of exposure rather than of pharmacology.

Machine learning has been brought to bear on this problem with notable success. Models trained on whole-cell growth-inhibition screens have identified antibacterial compounds that were then confirmed experimentally, including agents active against priority Gram-negative pathogens (Stokes et al., 2020; Liu et al., 2023), and the approach has since been extended to larger virtual libraries and to unconventional sequence space (Scalia et al., 2025; Torres et al., 2025). These results are genuine. They have also encouraged a broader expectation: that models can be trained to predict penetration and accumulation from chemical structure, and that such predictions can then guide compound selection in programmes targeting organisms and chemotypes the models have not seen.

This paper asks what such models can reasonably be expected to learn. Its argument is that the limit is not primarily one of algorithm or of sample size, but of what the available labels measure. Every label that can be generated at the scale machine learning requires (a minimum inhibitory concentration, a whole-cell accumulation value, an efflux ratio) is a composite in which influx across the outer membrane, active efflux, partitioning between cellular compartments and engagement with the target are summed into a single number. The separate contributions are not recoverable from that number. A model trained on it learns the composite, which is well defined only for the organisms, growth conditions and assay formats that produced the training data, and whose internal composition changes when any of those change.

The paper therefore sets out to identify the determinants that a structure-based model of Gram-negative penetration would have to capture; to analyse how assay design and dataset composition bound what published data can teach; to specify the circumstances in which predictions for novel chemotypes and unfamiliar species are unreliable; to develop a framework relating each determinant to the data, descriptors and validation design required to model it credibly; to state in falsifiable form when permeation prediction should and should not guide compound selection; and to propose reporting standards that would allow a reader to determine what a published model supports. The study is conceptual and generates no new measurements. Section 2 reviews the relevant literature, Section 3 states the design and verification protocol, Section 4 presents the framework, the propositions and the reporting standards and discusses their implications, including for antibacterial development intended for African health systems, and Sections 5 and 6 conclude.

2. LITERATURE REVIEW

2.1 CONCEPTUAL REVIEW

Penetration is not a single property. In the sense that matters for antibacterial activity it denotes the free concentration of a compound at its intracellular target, sustained over the period in which the target

© IIARJournal. All rights reserved.

must be occupied. No scalable assay measures that quantity. What is measured instead is one of a small family of proxies: a minimum inhibitory concentration, which reports the end point of the whole chain from medium to target engagement; a whole-cell accumulation value, usually by mass spectrometry of lysed cells, which reports total compound associated with the cell without distinguishing where it sits; a permeability coefficient for an isolated channel; or an efflux ratio comparing a wild-type strain with a pump-deficient mutant.

Each proxy is informative, and each collapses several processes into one number. This paper uses the term composite label for such a measurement, and treats the gap between the composite and the quantity of interest as the central conceptual problem rather than as measurement noise. The distinction matters because noise is reduced by more data, whereas the collapse of distinct processes into a single label is a property of the measurement that no quantity of further measurement of the same kind will undo.

2.2 THEORETICAL REVIEW

Four classes of determinant stand between a compound in the medium and a compound engaging an intracellular target in a Gram-negative cell. The first is transit of the outer membrane. Its asymmetric leaflet of lipopolysaccharide retards hydrophobic diffusion, so small hydrophilic compounds pass predominantly through water-filled porin channels whose constriction zones impose requirements on size, charge distribution and molecular shape rather than on bulk lipophilicity (Prajapati et al., 2021). Channel-level electrophysiology confirms that translocation depends on the alignment of a compound's dipole with the electric field inside the constriction, a property not captured by partition coefficients (Ghai, 2024).

The second is active efflux. In Gram-negative organisms the dominant systems belong to the resistance-nodulation-division superfamily, are driven by the proton-motive force across the inner membrane, and recruit accessory proteins to span both membranes (Zgurskaya, Walker, et al., 2021). They are strikingly polyspecific, exporting structurally unrelated compounds through overlapping binding sites, so whether a compound is a substrate is a matter of degree rather than of kind (Zgurskaya, Malloci, et al., 2021). Critically, efflux and the permeability barrier are not additive but synergistic: because efflux acts on whatever crosses, the steady-state intracellular concentration depends on the ratio of the two rates, and a modest change in either can produce a disproportionate change in accumulation (Zgurskaya, Walker, et al., 2021).

The third determinant is compartmental partitioning. The periplasm and the cytoplasm are distinct compartments with distinct access, and efflux from them proceeds at very different rates (Li et al., 2024). A whole-cell accumulation measurement sums them. The fourth is target engagement, which depends on the free rather than the total intracellular concentration and therefore on binding to intracellular macromolecules.

The theoretical claim advanced here follows from these four acting in series on a measurement that reports only their product. Influx, efflux, partitioning and engagement are not separately identifiable from a composite label, and the relative weight of each varies with the compound, the organism, the growth state and the assay. A model fitted to such labels therefore estimates a mixture whose composition is a property of the training setting rather than of chemistry alone, and its predictions are licensed only where that composition is preserved.

2.3 EMPIRICAL REVIEW

The empirical literature supports each element of this account. Systematic study of compound accumulation in Escherichia coli has yielded predictive guidance relating accumulation to the presence of an ionisable nitrogen, low three-dimensionality and limited flexibility, and that guidance has been used to convert Gram-positive-only agents into broad-spectrum ones (Muñoz & Hergenrother, 2021; Zhao et al., 2020). A standardised mass-spectrometric protocol with an accompanying predictive tool has made such measurements reproducible (Geddes et al., 2021).

The same literature demonstrates the limits of transfer. Accumulation in Pseudomonas aeruginosa was found to proceed largely independently of porins, so that the compound features governing entry differ from those established in E. coli, and a separate set of rules was required (Geddes et al., 2023). Predictors of efflux avoidance and inhibition in the same organism depended on species-specific descriptors, including interactions with particular residues of the MexB transporter (Mehla et al., 2021). Reviews of entry-facilitating strategies accordingly treat each priority pathogen's outer

© IIARJournal. All rights reserved.

membrane as a distinct problem (Saxena et al., 2023), and disrupting the barrier changes what the measured quantity means (MacNair & Brown, 2020).

Condition dependence is equally well evidenced. The contribution of efflux to intracellular accumulation has been shown to be large in actively growing cells and negligible in stationary phase, where envelope remodelling makes permeability the dominant term; the same strain and compound therefore yield different accumulation depending only on growth state (Whittle et al., 2021). Quantification of efflux in clinical isolates likewise shows that the contribution of expulsion to a resistance phenotype varies between strains carrying the same target mutations (Ferrand et al., 2023). Assay format contributes its own variation: vesicle-based permeation assays, channel electrophysiology and whole-cell accumulation each report different quantities (Richter et al., 2021).

Physics-based simulation offers a route around the composite label by computing permeation directly, through molecular dynamics of translocation (Deylami et al., 2024) or faster Brownian-dynamics approximations (Acharya et al., 2023), and by resolving how efflux and permeation jointly determine accumulation (Saha et al., 2020). These methods are mechanistic but compound- and channel-specific, and do not yet scale to the libraries that machine learning addresses.

Machine learning in this field has meanwhile succeeded where the composite label is exactly the quantity of interest. Models trained on whole-cell growth inhibition identified antibacterial compounds subsequently confirmed in the laboratory (Stokes et al., 2020), including an agent active against Acinetobacter baumannii that was discovered by screening against that single organism (Liu et al., 2023), and the approach has been scaled to large virtual libraries (Scalia et al., 2025) and extended to unconventional source material (Torres et al., 2025) and to mechanistic elucidation (Catacutan et al., 2025). It has also been applied to efflux-pump inhibitors (Bera et al., 2024) and resistance enzymes (Cheng et al., 2024), and generative approaches coupled to simulation have been used for antimicrobial design (Das et al., 2021). What these successes share is that the model was asked to rank compounds for the organism and assay on which it was trained.

The gap this paper addresses is therefore not a shortage of models or of mechanistic knowledge. It is the absence of an explicit account relating the determinants of penetration to the labels available for learning them, and of the validation designs that would license particular inferences from such models.

3. MATERIALS AND METHODS

3.1 RESEARCH DESIGN

This study is non-empirical and uses a conceptual research design. The unit of analysis is the published literature rather than compounds, bacteria or patients, and the output is a framework with falsifiable propositions rather than a dataset. No compound was synthesised, no accumulation was measured and no model was trained. There is consequently no sample, no randomisation and no statistical inference; what follows describes how the literature was assembled, how claims were admitted and on what basis the paper's two recurring judgements were made.

3.2 CORPUS CONSTRUCTION

Four bodies of literature were searched: outer-membrane permeation and porin-mediated uptake; efflux transport and its inhibition; compound accumulation measurement and the prediction tools built on it; and machine-learning discovery of antibacterial compounds. A supplementary search covered antimicrobial resistance burden and diagnostic capacity in sub-Saharan Africa, to anchor the analysis in the setting for which the authors write. Searches were conducted in PubMed and supplemented by citation tracing from the retrieved reviews.

Sources were included where they were peer-reviewed journal articles published between 2020 and 2026, treated at least one of these domains substantively, and reported enough methodological detail for the cited claim to be attributable, in particular, for permeation and accumulation studies, the organism, the growth conditions and the measured quantity. Preprints, conference abstracts, theses, commentaries without original synthesis, and sources whose claims could not be traced to the specific statement they were cited for were excluded. Priority was given to primary reports stating assay conditions explicitly, and to reviews that aggregate such conditions rather than restating conclusions.

© IIARJournal. All rights reserved.

3.3 CITATION VERIFICATION PROTOCOL

Every source admitted to the reference list was verified against its publisher record before citation, under a five-check protocol. First, the digital object identifier was resolved against the Crossref registry and the record compared field by field with the citation: authors and their order, title, journal, volume, issue, pagination or article number, and year. Second, each record was screened for retractions, corrections, errata and expressions of concern through the relation and update fields of the registry record and, independently, through the publication-type fields of the corresponding PubMed record. Third, the publication year recorded in the registry was adopted throughout, which for several items in this corpus is the year of advance online publication rather than that of the print issue to which the article was later assigned. Fourth, discrepancies between an indexing database and the publisher record were resolved in favour of the publisher record. Fifth, sources failing any check were excluded rather than corrected by inference.

No source is cited on the strength of an abstract alone, and no bibliographic field in the reference list was supplied from memory or inference. Where a cited study's organism, growth phase or assay format bears on the claim it supports, that detail is stated in the text rather than left implicit, because the argument of this paper is precisely that such details determine what a measurement means.

3.4 BASES OF JUDGEMENT

Two judgements recur and both are interpretive rather than measured. The first is whether a given measurement should be treated as a composite label or as a resolved one. A label is treated as composite where the reported quantity is produced by two or more of the four determinants defined in Section 2.2 acting in series without separate observation of each; as partially resolved where the design isolates one determinant by holding others fixed, as in comparisons between isogenic wild-type and efflux-deficient strains; and as resolved where the measurement reports a single physical process, as in single-channel electrophysiology or a computed permeation free-energy profile.

The second is which inference a reported validation licenses. A random split of a dataset is treated as licensing ranking within the chemotypes, organisms and conditions represented in that dataset; a scaffold- or chemotype-held-out split as licensing extrapolation to new chemical series within the same organism and condition; an organism-held-out design as licensing cross-species claims; and a condition-held-out design as licensing use under growth states or media absent from training. Where a study does not state its split design, the paper treats the licensed inference as undetermined rather than assuming the weakest or the strongest case.

3.5 DISCLOSURE OF ARTIFICIAL INTELLIGENCE USE

Generative artificial intelligence tools assisted with literature search, with retrieval and cross-checking of bibliographic records against the Crossref registry, and with language editing of the authors' drafted text. No text entered the manuscript without the authors' verification of its substantive claims against the cited sources. No such tool was used to generate data, to select citations without subsequent registry verification, or to formulate the framework and propositions in Section 4, which are the authors' own. The authors accept full responsibility for the content, including the accuracy of all references.

3.6 VALIDITY AND LIMITATIONS OF THE APPROACH

Construct validity concerns whether the four determinants correspond to processes that actually govern intracellular exposure; they are drawn from the permeation and efflux literature rather than invented. Descriptive validity concerns whether the composite-label classification fairly represents the cited studies; the criteria in Section 3.4 are stated so that a reader can contest a particular assignment. Theoretical validity concerns falsifiability; each proposition names the comparison that would contradict it. The principal limitation is that the analysis inherits the reporting practices of its corpus: where a study does not state growth phase, strain background or split design, the omission can be identified but not repaired, which is why Section 4.1.6 proposes reporting standards.

4. RESULTS AND DISCUSSION

4.1 RESULTS

4.1.1 What the available labels actually measure

Setting the measurements used in this field beside one another shows that they are not estimates of the same quantity at different precisions. They are estimates of different quantities: some collapse several of the four determinants into one number,

© IIARJournal. All rights reserved.

others report a single process, and each varies with a different set of experimental choices. Table 1 sets out the position.

Table 1. Measurements used in Gram-negative penetration research, classified by the criteria of Section 3.4, with the processes each collapses and the inference each licenses without further validation.

Measurement	Classification and processes collapsed	Principal source of between-study variation	Inference licensed without further validation
Minimum inhibitory concentration	Composite: influx, efflux, partitioning, target engagement and growth inhibition	Medium, inoculum, incubation time and strain background	Ranking of compounds within the organism and assay used
Whole-cell accumulation by mass spectrometry	Composite: influx, efflux and partitioning between periplasm and cytoplasm	Growth phase, wash protocol and the efflux background of the strain	Relative accumulation within the organism and growth state used
Efflux ratio across isogenic strains	Partially resolved: efflux relative to influx, under one growth state	Which transporters are deleted, and the growth state at measurement	Whether a compound is a substrate of the deleted transporters under those conditions
Outer-membrane permeability coefficient	Resolved: transit of a single barrier	Reconstitution system and lipid composition	Channel-level permeation for the channel and system studied
Single-channel electrophysiology	Resolved: translocation through one channel	Applied potential, buffer and lipid composition	Mechanistic interpretation confined to that channel
Computed permeation free energy	Resolved: transit of a defined barrier within a defined model	Force field, membrane model and sampling adequacy	A mechanistic hypothesis requiring experimental test

The fourth column is the one that matters for machine learning. A minimum inhibitory concentration supports the ranking of compounds within the organism and assay that produced it, and supports nothing else without further evidence; it cannot distinguish a compound that enters poorly from one that enters well and is expelled, or from one that reaches its target and binds weakly. Whole-cell accumulation removes target engagement from the composite but retains the others, and adds a dependence on the wash protocol and on the efflux background of the strain. Only the last three rows report single processes, and those are precisely the measurements that do not scale.

4.1.2 Non-identifiability and the consequences of mixture shift

The central analytical point follows directly. When several processes act in series and only their product is observed, the individual contributions are not identifiable: many different decompositions produce the same observed value. A model fitted to such observations does not estimate the determinants; it estimates whatever mixture of them generated the training data. This is not a defect of any particular algorithm, and it is not remedied by model capacity.

Two consequences follow. The first concerns interpretation. Feature-importance analyses performed on models trained on composite labels cannot support mechanistic conclusions, because a descriptor that appears important may be acting through influx in one part of the chemical space and through efflux avoidance in another. The second concerns generalisation. Prediction will degrade not in proportion to how chemically novel a compound is, but in proportion to how much the mixture has shifted, how much the relative contributions of influx, efflux and partitioning in the setting of use differ from those in training. The literature supplies a clean demonstration: accumulation in Pseudomonas aeruginosa proceeds largely without porin involvement, so the compound features that govern entry there differ from those established in Escherichia coli, and a separate set of rules was required (Geddes et al., 2023). The species are not distant chemically; the mixture is simply different.

4.1.3 A determinant-to-validation framework

The framework proposed here relates four layers that are usually discussed separately: the quantity of interest, the label available, the processes that separate them, and the validation design that would license a given use of the model.

Layer 1 states the quantity of interest, which for antibacterial purposes is the free concentration of compound at the intracellular target, sustained over the required period. Layer 2 names the label that will actually be used for training, and classifies it by Table 1 as composite, partially resolved or resolved. Layer 3 enumerates what separates the two: which processes the label collapses, and which experimental variables were held fixed in generating it. Layer 4 specifies the validation design, and with it the inference the model may be used to support. A random split licenses ranking within the training chemotypes, organism and conditions. A scaffold-held-out split licenses extrapolation to new chemical series within the same organism and condition. An organism-held-out design licenses cross-species use. A condition-held-out design licenses application to

© IIARJournal. All rights reserved.

growth states, media or infection sites absent from training.

The framework's operative rule is that the inference drawn from a model must not exceed the layer-4 design that was actually evaluated. This is a stronger constraint than it may appear. Most published permeation and antibacterial-activity models report random-split performance, which under this rule licenses only in-distribution ranking, which is, in fact, exactly how the most successful applications have used them, screening large libraries against the organism and assay on which the model was trained (Stokes et al., 2020; Liu et al., 2023; Scalia et al., 2025).

4.1.4 Three distinguishable modes of failure

The framework separates three failures that are often described together as poor generalisation. Chemotype failure occurs when a model is applied to a scaffold whose entry route differs from those in training, and is detected by scaffold-held-out evaluation. Species failure occurs when rules are transported to an organism with a different dominant uptake route; the porin-independent accumulation of Pseudomonas is the documented case (Geddes et al., 2023), and species-specific descriptors have been required even to predict efflux avoidance within a single organism (Mehla et al., 2021). Condition failure occurs when a model trained on data from one growth state is applied to another; that efflux contributes substantially to accumulation in growing cells and negligibly in stationary phase makes this a quantitative and not a hypothetical concern (Whittle et al., 2021), as does the two-order-of-magnitude difference between efflux rates from the periplasm and from the cytoplasm (Li et al., 2024).

These failures call for different remedies. Chemotype failure is addressed by broadening chemical coverage; species failure is addressed by resolved, organism-specific measurement, not by pooling species into a larger dataset; and condition failure is addressed by recording and stratifying on condition, which the present reporting practice of the field largely prevents.

4.1.5 Propositions

The framework yields six propositions, each stated so that a specified comparison could contradict it.

Proposition 1 (non-identifiability). Because influx, efflux, compartmental partitioning and target engagement are collapsed into every label that can be generated at the scale machine learning requires, their separate contributions are not identifiable from such data. A model trained on composite labels therefore cannot attribute its predictions to a mechanism, and mechanistic interpretations of feature importance in such models are unlicensed. Contradicted by a demonstration that the separate contributions can be recovered from composite labels alone, without additional resolved measurements.

Proposition 2 (mixture shift rather than chemical distance). Off-distribution prediction error is governed by the change in the relative weights of the collapsed processes between the training setting and the setting of use, not by chemical distance as such. Two compounds equally distant from the training set in descriptor space will therefore differ in prediction reliability if one preserves the training mixture and the other does not. Contradicted if prediction error tracks chemical distance once mixture composition is held constant.

Proposition 3 (species non-transfer is route-determined). Accumulation rules learned in one Gram-negative species fail to transfer to a species whose dominant route of entry differs, and the failure is predictable from the route rather than from phylogenetic proximity. Contradicted if transfer success correlates with phylogenetic distance after the dominant uptake route is controlled for.

Proposition 4 (condition non-exchangeability). Labels generated in actively growing and in stationary-phase cultures are not exchangeable, because the relative contribution of efflux to accumulation differs between them. Datasets pooling growth states therefore acquire structured label noise that no increase in sample size removes. Contradicted if models trained on pooled growth states match, on condition-held-out evaluation, models trained within a single growth state.

Proposition 5 (validation licenses the inference). The inference a permeation model licenses is fixed by its held-out design, not by its reported accuracy. Random-split performance is uninformative about scaffold-hopping, cross-species or cross-condition use, and a model reporting only random-split metrics supports only ranking within its training distribution. Contradicted if random-split performance predicts chemotype-held-out and organism-held-out performance across a range of models.

© IIARJournal. All rights reserved.

Proposition 6 (accuracy ceiling as label saturation). For a fixed composite label there exists an accuracy ceiling set by the variance of the unobserved decomposition across the dataset. Models approaching that ceiling are evidence of label saturation rather than of mechanistic understanding, and further gains require resolved measurements rather than larger composite datasets. Contradicted if accuracy continues to improve materially with dataset size after between-study condition variation has been accounted for.

4.1.6 Reporting standards

The propositions are testable only against a literature that records the conditions under which its labels were generated. Table 2 states the minimum that a study reporting a permeation- or accumulation-prediction model should disclose; each item corresponds to information whose absence, in assembling this paper's corpus, made it impossible to determine what a reported model would license.

Table 2. Reporting standards for permeation- and accumulation-prediction models in antibacterial discovery.

Reporting item	What should be stated
Organism and strain	The species and strain used, including whether the efflux background is wild type or pump-deficient
Growth state and medium	Growth phase at measurement, medium composition and incubation conditions
Measured quantity	What was measured, in what units, and which of the processes in Table 1 it collapses
Assay format	The assay used, and whether the quantity is composite, partially resolved or resolved
Compound set provenance	Where the compounds came from and which chemotypes are represented
Definition of inactives	How negative or non-accumulating examples were defined and obtained
Split design	Whether evaluation used random, scaffold-held-out, organism-held-out or condition-held-out splits
Performance by split	Performance reported separately for each split design rather than pooled
Applicability domain	The chemotypes, organisms and conditions within which predictions are claimed to hold
Prospective validation	Number of predicted compounds tested, number confirmed, and the resulting hit rate
Mechanistic claims	Whether any mechanistic interpretation is offered, and the resolved evidence supporting it

4.2 DISCUSSION

The prevailing treatment of Gram-negative penetration casts it as a hard but ordinary prediction problem: the determinants are complex, the data are scarce, and better models and larger datasets will close the gap. The analysis above suggests a different diagnosis. The binding constraint is not the model or the volume of data but the measurement. Every label available at scale reports a product of processes that cannot be recovered from it, and the composition of that product is a property of the organism, the growth state and the assay. Scaling a composite dataset improves estimation of the composite; it does not resolve the determinants, and it does not establish that a model will transfer.

This reading is supported, rather than contradicted, by the field's clearest success. The programme that produced usable accumulation guidance did not proceed by assembling heterogeneous published data; it generated a purpose-built dataset in a single organism under controlled conditions, derived compound-level rules from it, and tested those rules prospectively by converting Gram-positive-only agents into compounds active against Gram-negative bacteria (Muñoz & Hergenrother, 2021; Zhao et al., 2020), with a standardised protocol subsequently published so the measurement could be reproduced (Geddes et al., 2021). That is resolution by experimental design rather than by inference from aggregated labels, and its limits appeared exactly where the framework predicts: in another species with another dominant route of entry (Geddes et al., 2023).

The two methodological traditions in this field have complementary deficits. Physics-based approaches resolve the determinants, a free-energy profile for translocation through a named channel is a statement about one process (Deylami et al., 2024), and faster approximations extend the same logic (Acharya et al., 2023), but they do not scale to the libraries that discovery programmes screen. Machine learning scales but does not resolve. The productive relationship between them is therefore not competition but division of labour: resolved measurement and simulation are the appropriate instruments for establishing where a scaled model's predictions apply, and for detecting when the mixture has shifted. On this view the most valuable contribution of mechanistic work to machine

© IIARJournal. All rights reserved.

learning is not better features but a defensible applicability domain.

It should be said plainly that the successes of machine learning in antibacterial discovery are genuine and that nothing here diminishes them. Whole-cell growth inhibition is a composite label, but where the object is to find a compound that inhibits growth of a particular organism, the composite is the quantity of interest, and a model trained on it is being asked the question it can answer. The difficulty arises only when such a model is asked a different question, about a mechanism, a new species, or a growth state it has not seen, and the answer is read with the confidence earned on the first.

4.2.1 Implications for antibacterial development intended for African health systems

The composition of training data is not a neutral technical matter when models are used to guide development for particular settings. Resistance in sub-Saharan Africa is driven substantially by Gram-negative organisms, and the burden is projected to rise sharply without intervention (Totaro et al., 2025); carbapenem resistance among Gram-negative pathogens has risen faster than resistance to any other class globally (Naghavi et al., 2024). The strains in which permeation and accumulation are characterised, however, are predominantly reference laboratory strains, and strain background is precisely what determines the efflux contribution to a given phenotype, quantitative study of clinical isolates shows the contribution of expulsion varying between strains carrying comparable target mutations (Ferrand et al., 2023).

Two consequences follow for the region. First, the diagnostic gap that limits routine identification and susceptibility testing (Hermans et al., 2024) also limits the generation of local susceptibility data, so circulating isolates are under-represented in exactly the datasets from which models are built; a model's applicability domain, stated honestly, may not include the organisms that matter most locally. Second, because the framework holds that organism and strain background determine the mixture, this under-representation is not corrected by enlarging datasets drawn from the same reference strains. It is corrected only by measurement in locally circulating isolates. The practical recommendation is therefore a representativeness constraint: a model intended to guide development for a given setting should state the strain provenance of its training data, and prospective validation in locally circulating isolates should be treated as a precondition of use rather than as confirmatory work.

4.2.2 Limitations and directions for further work

The framework is conceptual and its propositions are untested; it earns its place only if the comparisons named in Section 4.1.5 are attempted. The most exposed is Proposition 2, which asserts that mixture shift rather than chemical distance governs transfer; testing it requires a dataset in which both quantities vary independently, which does not presently exist and would have to be built. The analysis is also bounded by the reporting practices of its corpus: where studies do not record growth phase, strain background or split design, the omission can be identified but not repaired, which is why Table 2 exists. Finally, the paper addresses penetration and says nothing about target engagement kinetics or resistance emergence, each of which imposes its own constraints on whether an accumulating compound becomes a useful drug.

5. CONCLUSION AND RECOMMENDATION

Machine learning can be expected to learn what its labels contain. In Gram-negative antibacterial discovery those labels are composites in which influx, efflux, compartmental partitioning and target engagement are summed into a single number, and the relative weight of each term varies with the organism, the growth state and the assay. A model fitted to such data estimates a mixture, not a mechanism, and its predictions are licensed only where that mixture is preserved. This is a limit of measurement rather than of method, and it is not relieved by larger datasets of the same kind.

The framework proposed here makes the limit explicit by relating the quantity of interest to the available label, the processes that separate them, and the validation design that licenses a given use. It distinguishes chemotype, species and condition failure as separate modes with separate remedies, states six falsifiable propositions, and sets out reporting standards that would allow a reader to determine what a published model supports.

Four recommendations follow. Investigators building permeation or accumulation models should state the processes their labels collapse and should report performance separately for random, scaffold-held-out, organism-held-out and condition-held-out splits, since these license different inferences. The field should treat condition-annotated and organism-resolved measurement as a higher priority than further accumulation of heterogeneous composite

© IIARJournal. All rights reserved.

data. Journals and reviewers should expect the disclosures in Table 2, and in particular should not accept mechanistic interpretations of models trained on composite labels without resolved supporting evidence. And for development intended for African health systems, model applicability should be established in locally circulating isolates rather than assumed from reference strains.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

This paper makes four contributions. First, it introduces the composite label as an analytical category for this field and classifies the measurements in routine use by what each collapses and what each licenses. Second, it advances a non-identifiability argument showing why the limit on machine learning here is one of measurement rather than of algorithm or sample size, and derives from it the claim that transfer is governed by mixture shift rather than by chemical distance. Third, it offers a four-layer framework linking the quantity of interest to label, confound and validation design, with an operative rule constraining the inference a model may support, and distinguishes three modes of generalisation failure with distinct remedies. Fourth, it supplies reporting standards for permeation-prediction models and extends the analysis with a representativeness constraint for settings, including much of sub-Saharan Africa, whose circulating pathogens are under-represented in the data from which such models are built.

7. REFERENCES

Acharya, A., Jana, K., Gurvic, D., Zachariae, U., & Kleinekathöfer, U. (2023). Fast prediction of antibiotic permeability through membrane channels using Brownian dynamics. *Biophysical Journal*, *122*(14), 2996–3007. https://doi.org/10.1016/j.bpj.2023.03.035

Bera, A., Roy, R. K., Joshi, P., & Patra, N. (2024). Machine learning-guided discovery of AcrB and MexB efflux pump inhibitors. *The Journal of Physical Chemistry B*, *128*(3), 648–663. https://doi.org/10.1021/acs.jpcb.3c05845

Catacutan, D. B., Tran, V., Arnold, A., Alexander, J., Corso, G., Yousefi, Y., Tu, M. M., McLellan, S., Tertigas, D., Corneil, K., Magolan, J., Surette, M. G., Brown, E. D., Coombes, B. K., Barzilay, R., & Stokes, J. M. (2025). Discovery and artificial intelligence-guided mechanistic elucidation of a narrow-spectrum antibiotic. *Nature Microbiology*, *10*(11), 2808–2822. https://doi.org/10.1038/s41564-025-02142-0

Cheng, Z., Aitha, M., Thomas, C. A., Sturgill, A., Fairweather, M., Hu, A., Bethel, C. R., Rivera, D. D., Dranchak, P., Thomas, P. W., Li, H., Feng, Q., Tao, K., Song, M., Sun, N., Wang, S., Silwal, S. B., Page, R. C., Fast, W., . . . Crowder, M. W. (2024). Machine learning models identify inhibitors of New Delhi metallo-β-lactamase. *Journal of Chemical Information and Modeling*, *64*(10), 3977–3991. https://doi.org/10.1021/acs.jcim.3c02015

Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., dos Santos, C., Chen, P.-Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., & Mojsilovic, A. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, *5*(6), 613–623. https://doi.org/10.1038/s41551-021-00689-x

Deylami, J., Chng, S. S., & Yong, E. H. (2024). Elucidating antibiotic permeation through the Escherichia coli outer membrane: Insights from molecular dynamics. *Journal of Chemical Information and Modeling*, *64*(21), 8310–8321. https://doi.org/10.1021/acs.jcim.4c01249

Ferrand, A., Vergalli, J., Bosi, C., Pantel, A., Pagès, J.-M., & Davin-Regli, A. (2023). Contribution of efflux and mutations in fluoroquinolone susceptibility in MDR enterobacterial isolates: A quantitative and molecular study. *Journal of Antimicrobial Chemotherapy*, *78*(6), 1532–1542. https://doi.org/10.1093/jac/dkad122

Geddes, E. J., Gugger, M. K., Garcia, A., Chavez, M. G., Lee, M. R., Perlmutter, S. J., Bieniossek, C., Guasch, L., & Hergenrother, P. J. (2023). Porin-independent accumulation in Pseudomonas enables antibiotic discovery. *Nature*, *624*(7990), 145–153. https://doi.org/10.1038/s41586-023-06760-8

Geddes, E. J., Li, Z., & Hergenrother, P. J. (2021). An LC-MS/MS assay and complementary web-based tool to quantify and predict compound accumulation in E. coli. *Nature Protocols*, *16*(10), 4833–4854. https://doi.org/10.1038/s41596-021-00598-y

Ghai, I. (2024). Electrophysiological insights into antibiotic translocation and resistance: The impact of outer membrane proteins. *Membranes*,

© IIARJournal. All rights reserved.

14(7), 161. https://doi.org/10.3390/membranes14070161

Hermans, L. E., Centner, C. M., Morel, C. M., Mbamalu, O., Bonaconsa, C., Ferreyra, C., Lindahl, O., & Mendelson, M. (2024). Point-of-care diagnostics for infection and antimicrobial resistance in sub-Saharan Africa: A narrative review. *International Journal of Infectious Diseases*, *142*, Article 106907. https://doi.org/10.1016/j.ijid.2023.11.027

Li, Y., Wilhelm, M. J., Wu, T., Hu, X.-H., Ruiz, O. N., & Dai, H.-L. (2024). Quantifying bacterial efflux within subcellular domains of Pseudomonas aeruginosa. *Applied and Environmental Microbiology*, *90*(11), Article e01447-24. https://doi.org/10.1128/aem.01447-24

Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., Chiappino-Pepe, A., Syed, S. A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M. G., Coombes, B. K., Jaakkola, T., Barzilay, R., Collins, J. J., & Stokes, J. M. (2023). Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nature Chemical Biology*, *19*(11), 1342–1350. https://doi.org/10.1038/s41589-023-01349-8

MacNair, C. R., & Brown, E. D. (2020). Outer membrane disruption overcomes intrinsic, acquired, and spontaneous antibiotic resistance. *mBio*, *11*(5), Article e01615-20. https://doi.org/10.1128/mBio.01615-20

Mehla, J., Malloci, G., Mansbach, R., López, C. A., Tsivkovski, R., Haynes, K., Leus, I. V., Grindstaff, S. B., Cascella, R. H., D'Cunha, N., Herndon, L., Hengartner, N. W., Margiotta, E., Atzori, A., Vargiu, A. V., Manrique, P. D., Walker, J. K., Lomovskaya, O., Ruggerone, P., . . . Zgurskaya, H. I. (2021). Predictive rules of efflux inhibition and avoidance in Pseudomonas aeruginosa. *mBio*, *12*(1), Article e02785-20. https://doi.org/10.1128/mBio.02785-20

Muñoz, K. A., & Hergenrother, P. J. (2021). Facilitating compound entry as a means to discover antibiotics for Gram-negative bacteria. *Accounts of Chemical Research*, *54*(6), 1322–1333. https://doi.org/10.1021/acs.accounts.0c00895

Naghavi, M., Vollset, S. E., Ikuta, K. S., Swetschinski, L. R., Gray, A. P., Wool, E. E., Robles Aguilar, G., Mestrovic, T., Smith, G., Han, C., Hsu, R. L., Chalek, J., Araki, D. T., Chung, E., Raggi, C., Gershberg Hayoon, A., Davis Weaver, N., Lindstedt, P. A., Smith, A. E., . . . Murray, C. J. L. (2024). Global burden of bacterial antimicrobial resistance 1990–2021: A systematic analysis with forecasts to 2050. *The Lancet*, *404*(10459), 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1

Prajapati, J. D., Kleinekathöfer, U., & Winterhalter, M. (2021). How to enter a bacterium: Bacterial porins and the permeation of antibiotics. *Chemical Reviews*, *121*(9), 5158–5192. https://doi.org/10.1021/acs.chemrev.0c01213

Richter, R., Kamal, M. A. M., Koch, M., Niebuur, B.-J., Huber, A.-L., Goes, A., Volz, C., Vergalli, J., Kraus, T., Müller, R., Schneider-Daum, N., Fuhrmann, G., Pagès, J.-M., & Lehr, C.-M. (2021). An outer membrane vesicle-based permeation assay (OMPA) for assessing bacterial bioavailability. *Advanced Healthcare Materials*, *11*(5), Article 2101180. https://doi.org/10.1002/adhm.202101180

Saha, P., Sikdar, S., Krishnamoorthy, G., Zgurskaya, H. I., & Rybenkov, V. V. (2020). Drug permeation against efflux by two transporters. *ACS Infectious Diseases*, *6*(4), 747–758. https://doi.org/10.1021/acsinfecdis.9b00510

Saxena, D., Maitra, R., Bormon, R., Czekanska, M., Meiers, J., Titz, A., Verma, S., & Chopra, S. (2023). Tackling the outer membrane: Facilitating compound entry into Gram-negative bacterial pathogens. *npj Antimicrobials and Resistance*, *1*(1), Article 17. https://doi.org/10.1038/s44259-023-00016-1

Scalia, G., Rutherford, S. T., Lu, Z., Buchholz, K. R., Skelton, N., Chuang, K., Diamant, N., Hütter, J.-C., Luescher, J.-M., Miu, A., Blaney, J., Gendelev, L., Skippington, E., Zynda, G., Dickson, N., Koziarski, M., Bengio, Y., Regev, A., Tan, M.-W., & Biancalani, T. (2025). Deep-learning-based virtual screening of antibacterial compounds. *Nature Biotechnology*, *44*(7), 1154–1167. https://doi.org/10.1038/s41587-025-02814-6

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S.,

© IIARJournal. All rights reserved.

Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

Torres, M. D. T., Wan, F., & de la Fuente-Nunez, C. (2025). Deep learning reveals antibiotics in the archaeal proteome. *Nature Microbiology*, *10*(9), 2153–2167. https://doi.org/10.1038/s41564-025-02061-0

Totaro, V., Guido, G., Cotugno, S., De Vita, E., Asaduzzaman, M., Patti, G., Segala, F. V., Putoto, G., Frallonardo, L., Farkas, F. B., Lakatos, B., Veronese, N., Locantore, P., Di Gennaro, F., & Saracino, A. (2025). Antimicrobial resistance in sub-Saharan Africa: A comprehensive landscape review. *The American Journal of Tropical Medicine and Hygiene*, *113*(2), 253–263. https://doi.org/10.4269/ajtmh.25-0035

Whittle, E. E., McNeil, H. E., Trampari, E., Webber, M., Overton, T. W., & Blair, J. M. A. (2021). Efflux impacts intracellular accumulation only in actively growing bacterial cells. *mBio*, *12*(5), Article e02608-21. https://doi.org/10.1128/mBio.02608-21

Zgurskaya, H. I., Malloci, G., Chandar, B., Vargiu, A. V., & Ruggerone, P. (2021). Bacterial efflux transporters' polyspecificity – a gift and a curse?. *Current Opinion in Microbiology*, *61*, 115–123. https://doi.org/10.1016/j.mib.2021.03.009

Zgurskaya, H. I., Walker, J. K., Parks, J. M., & Rybenkov, V. V. (2021). Multidrug efflux pumps and the two-faced Janus of substrates and inhibitors. *Accounts of Chemical Research*, *54*(4), 930–939. https://doi.org/10.1021/acs.accounts.0c00843

Zhao, S., Adamiak, J. W., Bonifay, V., Mehla, J., Zgurskaya, H. I., & Tan, D. S. (2020). Defining new chemical space for drug penetration into Gram-negative bacteria. *Nature Chemical Biology*, *16*(12), 1293–1302. https://doi.org/10.1038/s41589-020-00674-6

© IIARJournal. All rights reserved.