Journal of Advanced Pharmaceutical Sciences and Natural Products

Journal of Advanced Pharmaceutical Sciences and Natural Products (JAPSNP)

ARTIFICIAL INTELLIGENCE FOR PROPERTY PREDICTION IN DRUG PRODUCT FORMULATION: QSPR MODELS, FEATURE ENGINEERING, AND VALIDATION

Jyoti Pandit*

Assistant Professor, Jigyasa University, Dehradun, Uttarakhand, India

Corresponding author:
Jyoti Pandit

Email: jyotipandit082002@gmail.com

ABSTRACT

The formulation of a drug product depends on a dense web of physicochemical and performance properties, from aqueous solubility and dissolution to the physical stability of an amorphous dispersion and the particle size of a nanosuspension. Measuring each property empirically for every candidate and excipient combination is slow and consumes scarce active material, and this cost is what has drawn quantitative structure-property relationship modelling and, more recently, machine learning into formulation science. This review examines how artificial intelligence is now used to predict formulation-relevant properties, and it argues that the discipline is limited less by the choice of learning algorithm than by three underdeveloped foundations: the representation of the formulation rather than the molecule alone, the scale and quality of the underlying data, and the rigour of validation. The prediction targets are surveyed by property class, spanning solubility, amorphous solid dispersion stability, tablet mechanical behaviour, nanoparticle attributes, and permeability, with the reported models and their measured performance placed side by side. The representation problem is then treated directly, comparing expert descriptors, structural fingerprints, and learned graph or sequence representations, and noting where each fails when the input is a mixture and a process rather than a single structure. The validation section sets out what a trustworthy model must demonstrate, drawing on the established principles for structure-property models: a defined applicability domain, honest external testing, and metrics that reflect real predictivity rather than fitted correlation. The synthesis is that formulation artificial intelligence will become decision-grade only when shared, standardised datasets and applicability-domain-aware, externally validated, interpretable models replace the single-laboratory, internally validated demonstrations that dominate the current literature.

Keywords: quantitative structure-property relationship; machine learning; drug formulation; molecular descriptors; model validation; applicability domain; amorphous solid dispersion; deep learning.

1. Introduction

A drug substance becomes a medicine only after it is formulated into a product that dissolves, remains stable, and delivers a reproducible dose. The decisions that produce such a product, which polymer to disperse a poorly soluble drug in, how much surfactant a nanosuspension needs, whether a tablet will fracture under compression, rest on properties that have traditionally been established by experiment. Each experiment costs time and material, and the combinatorial space of drugs, excipients, ratios, and process settings is far too large to explore exhaustively at the bench. The appeal of a predictive model is therefore immediate: if a property can be estimated reliably from the structure of the drug and the composition of the formulation, the experimental effort can be concentrated on the candidates most likely to succeed.[1]

Predicting a property from chemical structure is not a new ambition. The quantitative structure-activity and structure-property relationship, in which a measurable endpoint is expressed as a mathematical function of structural descriptors, was formalised in the 1960s through the linear free-energy approach of Hansch and Fujita and the additive substituent model of Free and Wilson.[2,3] Those methods were built for small, congeneric series and for a handful of interpretable descriptors fitted by linear regression. The modern incarnation retains the same logic, a function mapping structure to property, but replaces the linear model with flexible learning algorithms and the handful of descriptors with hundreds or thousands of features, and it extends the target from biological activity to the physicochemical and manufacturing properties that govern a drug product.[4] This continuity matters, because the hard-won lessons of classical structure-property modelling about overfitting, chance correlation, and the limits of extrapolation apply with equal or greater force to the data-hungry models now in use.

Interest in applying artificial intelligence to pharmaceutical development has grown quickly, spanning discovery, formulation optimisation, process control, and quality assurance.[5,6] Regulators have moved in step. The United States Food and Drug Administration issued a discussion paper on the use of artificial intelligence and machine learning in drug development in 2023 and followed it with draft guidance proposing a risk-based credibility assessment framework for models used to support regulatory decisions.[7,8] The European Medicines Agency set out its current thinking across the medicine lifecycle in a reflection paper adopted in 2024.[9] These documents share a premise that is central to this review: a prediction is only useful for a regulated product if its reliability can be characterised and defended, which places the burden on validation rather than on predictive novelty.

This review concentrates on property prediction for the drug product, the formulated dosage form, rather than on target-based activity prediction for drug discovery, which is reviewed extensively elsewhere. Its scope is the modelling of formulation-relevant physicochemical and

performance properties, the engineering of the features that such models consume, and the methodology by which their predictions are validated. The argument advanced throughout is that the current bottleneck is not algorithmic. Powerful learning methods are freely available and are routinely applied. The bottleneck is threefold: the field still struggles to represent a formulation, which is a mixture processed under specified conditions, rather than an isolated molecule; the datasets that underpin most published models are small, heterogeneous, and rarely shared; and validation practice frequently reports fitted performance in place of genuine external predictivity within a stated domain. Sections 3 to 5 examine the prediction targets, the representation problem, and the validation problem in turn, and the discussion sets out what would move the field from demonstration to dependable use.

2. Scope and evidence base

The literature synthesised here was identified through structured searches of biomedical and cheminformatics databases for machine learning and quantitative structure-property relationship studies addressing formulation-relevant properties, together with the foundational methodological literature on molecular representation and model validation and the current regulatory guidance. Primary studies were selected to represent the range of property classes and modelling strategies rather than to enumerate every report, consistent with a narrative rather than a systematic review. Priority was given to studies that reported the dataset size, the features used, and a quantitative measure of predictive performance, because these are the elements on which the review's argument about validation depends. Representative studies are summarised in Table 1, the analytical workflow common to the field is shown in Figure 1, and reported predictive performance across the surveyed studies is compared in Figure 2.

Table 1. Representative machine learning and quantitative structure-property relationship studies for the prediction of drug product formulation properties. Values are the headline figures reported by each study under its own dataset and validation protocol.

Property target	Study	Model (best)	Principal features	Data size	Reported performance
Aqueous solubility	Delaney 2004 [10]	Linear regression (ESOL)	4 physicochemical descriptors	2,874 train / 528 test	Test R^2 0.55; average absolute error 0.83 log
Aqueous solubility (reference set)	Sorkun 2019 [11]	Curated benchmark dataset	RDKit 2D descriptors	9,982 compounds	Reliability-graded standard set
Aqueous solubility	Zheng 2024 [12]	Graph network vs descriptor models	>4,000 descriptors or learned graphs	AqSolDB-derived	Graphs best on clean data; descriptors more robust
Solubility in	Bao 2024	Gradient-	Solute and	~27,000	Mean absolute

binary solvents	[13]	boosted trees	solvent descriptors, temperature	measurements	error ~0.33 log units
Amorphous dispersion physical stability	Han 2019 [14]	Random forest (of 8)	>20 molecular descriptors	646 data points	Accuracy 82.5% (dissimilarity split)
Hot-melt-extrusion amorphization	Jiang 2023 [15]	LightGBM + ECFP	Material attributes and process parameters	760 formulations	Accuracy 92.8%; AUC 0.932
Hot-melt-extrusion chemical stability	Jiang 2023 [15]	XGBoost + ECFP	Material attributes and process parameters	495 formulations	Accuracy 96.0%; AUC 0.944
Tablet breaking force and disintegration	Akseli 2017 [17]	ML with ultrasonic elastic measurement	Composition, process, elastic properties	Small-scale	Validated predictive framework
Tablet disintegration time	Ghazwani 2025 [18]	Sparse Bayesian learning	Molecular, material, formulation variables	~2,000 data points	Test R^2 0.960; RMSE 10.07
Lipid nanoparticle mRNA potency	Xu 2024 [19]	Graph network + descriptor (AGILE)	Lipid graph and Mordred descriptors	1,200 lipids screened	Predicted-observed r 0.78; 7.8-fold vs benchmark
Lipid nanoparticle transfection	Chan 2025 [20]	Transformer (COMET)	Structure, composition, formulation parameters	>3,000 LNPs / >6,000 points	Test r 0.866; 79.6% top-hit accuracy
Liposome size and polydispersity	Hoseini 2023 [21]	Boosting ensemble	Sonication time, temperature, lipids	Experimental set	Size mean absolute error ~1.65 nm
Drug nanocrystal size and polydispersity	He 2020 [22]	LightGBM	Process variables by preparation method	910 size / 341 PDI points	Strong for milling and homogenisation; weak for precipitation
Caco-2 permeability	Acuña 2024 [23]	SVM-RF-GBM ensemble	41 descriptors selected from >7,000	1,817 compounds	Test R^2 0.76; RMSE 0.38
Spray-dried particle size	Muñoz 2024 [24]	Tensor decomposition + PLS	11 process parameters and material attributes	216+139 / 68+51 batches	R^2 0.90 and 0.969
Nanosuspension stability and diameter	Zulbeari 2025 [25]	XGBoost (of 8) + SHAP	Stabiliser type and concentration, bead size	72 data points	Accuracy 0.91; F1 0.91

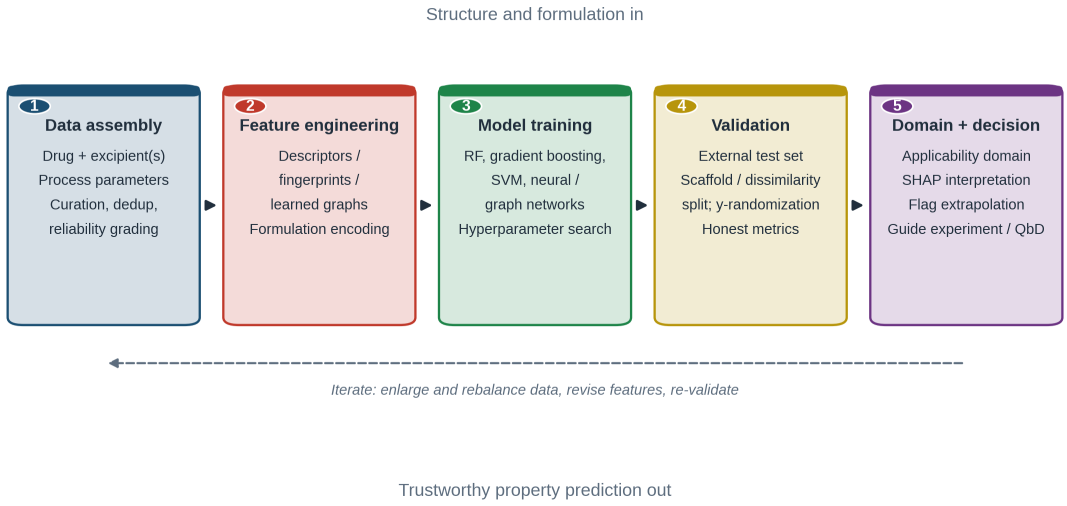

Figure 1. A workflow for trustworthy property prediction in drug product formulation. The pipeline proceeds from assembly and curation of paired drug, excipient and process data (1), through feature engineering that must encode the formulation and not the drug molecule alone (2), model training (3), and validation on an external, structure-aware split (4), to definition of the applicability domain, interpretation, and a development decision (5). The dashed return path denotes the iterative enlargement and rebalancing of data and revision of features that the current small-data regime requires. RF, random forest; SVM, support vector machine; SHAP, Shapley additive explanations; QbD, quality by design.

3. What is being predicted: property targets in formulation

3.1 Solubility and dissolution

Aqueous solubility is the property most often modelled, both because it constrains oral absorption and because comparatively large public datasets exist. The classical benchmark remains the ESOL model of Delaney, which estimated intrinsic aqueous solubility from four interpretable descriptors, calculated logarithm of the partition coefficient, molecular weight, rotatable bond count, and aromatic proportion, and reported a correlation coefficient of 0.72 on training data and 0.55 on a 528-compound blind test.[10] The gap between those two figures is itself instructive, and it recurs throughout the field. Progress since has come as much from data curation as from algorithms, exemplified by AqSolDB, a reference set of 9,982 unique compounds assembled from nine sources with reliability grading, which has become a standard training and benchmarking resource.[11]

Contemporary work applies graph neural networks and descriptor-based ensembles to these datasets. A systematic reassessment found that graph-based representations achieved the best predictive accuracy on high-quality data but were more sensitive to label noise, whereas descriptor-based models were more robust and more interpretable, a trade-off that has practical consequences when formulation data are noisy.[12] The solubility problem also generalises usefully to the formulation context when the solvent is not water alone. A model trained on roughly 27,000 solubility measurements across 123 solutes and 110 binary solvent mixtures reached a mean absolute error near 0.33 logarithm units

with gradient-boosted trees, yet its prospective error grew for compounds lying outside the chemical space of the training set.[13] That degradation on out-of-distribution inputs is a direct manifestation of the applicability-domain problem developed in Section 5.

3.2 Amorphous solid dispersions and stability

Rendering a poorly soluble drug amorphous within a polymer can raise its apparent solubility, but the amorphous state is thermodynamically metastable and may recrystallise, so physical stability over shelf life is the property that decides whether such a formulation is viable. Predicting stability is harder than predicting solubility because it depends on drug-polymer interactions, drug loading, and processing, and because stability datasets are small and often imbalanced toward the stable class. An early and widely cited study compared eight algorithms on 646 stability data points described by more than twenty molecular descriptors, and reported a random forest as the best model with a prediction accuracy of 82.5% under a maximum-dissimilarity data split.[14] The choice of a dissimilarity-based split rather than a random one is notable, because it makes the test harder and the reported accuracy more credible.

Later work broadened the inputs to include the manufacturing process. A study of hot-melt extruded dispersions assembled 760 formulations for the amorphisation endpoint and 495 for chemical stability, drawn from 49 active ingredients and 38 excipients across 158 publications, and combined critical material attributes with critical process parameters such as barrel temperature and screw speed.[15] Using extended-connectivity fingerprints with gradient boosting, the amorphisation model reached 92.8% accuracy and the chemical stability model 96.0%.[15] These figures are high, and they illustrate both the promise of the approach and the caution it warrants: accuracies near the ceiling on modest, literature-assembled datasets can reflect genuine signal, but they can also reflect redundancy between training and test compounds, which is precisely why the validation discipline of Section 5 is not optional. The class-imbalance problem that afflicts stability data has itself been addressed with resampling and dimensionality reduction inside deep learning pipelines, an approach that improves handling of the rare unstable class but does not remove the underlying scarcity of data.[16]

3.3 Tablet and compaction properties

The mechanical and disintegration behaviour of a tablet depends on the interaction of formulation composition with the compaction process, and these relationships have long resisted first-principles prediction. Machine learning has been applied to bridge the gap, in some cases coupled with rapid physical measurement. A practical framework combined machine learning with non-destructive ultrasonic measurement of elastic properties to predict tablet breaking force and disintegration time from small quantities of material, offering a route to rational prediction early in development.[17] More recent work has modelled disintegration time directly from formulation and material properties, comparing

Bayesian regression methods and reporting a test correlation coefficient of 0.960 with a sparse Bayesian learning model on roughly 2,000 data points.[18] Tablet properties are a useful reminder that a formulation model must ingest process and composition variables, not molecular structure alone, and that the most informative features are frequently the physical and formulation parameters rather than the descriptors of the active ingredient.

3.4 Nanoparticles and complex formulations

Complex and nanoparticulate products present the richest and most active application area, because their critical attributes, particle size, polydispersity, encapsulation, and transfection efficiency, depend on many interacting composition and process variables that defeat intuition. Lipid nanoparticles for nucleic acid delivery are the leading example. A deep learning platform for ionizable lipid design coupled a graph neural network encoder with a molecular descriptor encoder, screened a virtual library of roughly 60,000 structures, and identified a lead lipid that delivered messenger RNA to muscle 7.8-fold more effectively than an established benchmark lipid, with a predicted-to-observed correlation of 0.78 on intramuscular delivery.[19] A transformer-based model trained on a library of more than 3,000 lipid nanoparticles and over 6,000 labelled measurements predicted transfection with a test-set correlation above 0.86 and identified in vivo hits that outperformed clinical benchmark lipids by large margins.[20] These studies are among the most convincing in the field, in part because they close the loop with prospective experimental validation rather than resting on a held-out split.

Simpler nanoscale systems have been modelled similarly. An ensemble boosting model predicted liposome size and polydispersity index from sonication time, extrusion temperature, and lipid composition, reporting a mean absolute error of about 1.65 nm for size.[21] A model for drug nanocrystals trained on 910 size and 341 polydispersity data points performed well for milling and homogenisation but poorly for antisolvent precipitation, a failure the authors attributed candidly to poor reproducibility of that process in the source data.[22] That result is valuable precisely because it shows a model inheriting the noise of its data, and because it was reported rather than concealed.

3.5 Other product-relevant properties

Beyond these principal targets, machine learning has been applied to permeability, process yield, and suspension stability, each relevant to formulation decisions. An ensemble model predicted Caco-2 apparent permeability with a test correlation of 0.76 using 41 descriptors selected from more than 7,000, illustrating both the reach of feature selection and its necessity when descriptor spaces are vast.[23] Spray drying, a common process for producing dispersion intermediates, has been modelled with tensor decomposition and partial least squares to predict median particle size with correlation coefficients of 0.90 and above.[24] Nanosuspension physical stability and median diameter have been predicted from stabiliser type and concentration and milling bead size, with a gradient boosting

model reaching 0.91 accuracy and classification, and with feature attribution identifying stabiliser concentration as the dominant factor.[35] Across these varied targets a consistent pattern emerges: the model is only as good as the representation of the formulation it is given, which is the subject of the next section.

4. Feature engineering and molecular representation

If a single theme separates competent formulation modelling from the rest, it is how the drug and the formulation are turned into numbers. The learning algorithm operates on this representation, and no algorithm can recover information that the representation has discarded.

4.1 Expert descriptors

The oldest and still most widely used representation is the molecular descriptor, a numerical value computed from structure that encodes a constitutional, topological, electronic, or geometric feature of the molecule.[26] Thousands of such descriptors have been defined, and open software now computes them at scale. Mordred calculates more than 1,800 two- and three-dimensional descriptors and was designed to overcome the licensing and maintenance limits of earlier tools.[27] PaDEL computes 797 descriptors and ten fingerprint types on the Chemistry Development Kit framework, providing a free and reproducible alternative to commercial engines.[28] Descriptors are interpretable, which is their principal virtue: a model built on them can often be related back to a physical mechanism, which matters for regulatory acceptance and for scientific trust. Their weakness is that the informative descriptors for a given property are not known in advance, so the modeller computes many and must then select among them, a step that introduces its own risks discussed below.

4.2 Structural fingerprints

A fingerprint encodes the presence of substructural patterns as a bit vector. Two families dominate. Dictionary-based keys such as the reoptimised MACCS set use a fixed list of 166 predefined substructures, each bit signalling whether a pattern is present.[29] Circular fingerprints, of which extended-connectivity fingerprints are the standard, are generated by iteratively hashing the neighbourhood around each atom out to a chosen diameter, and were designed specifically for structure-activity and structure-property modelling rather than for substructure search.[30] Fingerprints are fast, require no descriptor selection, and capture local structure well, which is why the hot-melt extrusion stability models above achieved their best performance with extended-connectivity fingerprints.[15] They are, however, sparse and less directly interpretable than descriptors, and a bit collision can place unrelated substructures in the same feature.

4.3 Learned representations

The most significant methodological shift of the past decade is the move from fixed representations to representations learned from the data. Message passing neural networks treat the molecule as a graph and learn atom and bond

features by propagating information across the graph, so the representation is optimised for the property being predicted rather than fixed in advance.[31] A directed message passing architecture that augments the learned graph representation with global molecular features has become a widely used reference implementation and is competitive with or superior to fixed fingerprints across many datasets.[32] A parallel line treats the simplified molecular-input line-entry string as a language and pretrains transformer models on large unlabelled corpora, transferring the learned representation to property prediction with performance that scales with the pretraining set.[33] Learned representations are powerful when data are abundant, but they are data-hungry and opaque, and both properties are liabilities in formulation science, where data are scarce and interpretability is prized.

4.4 Representing the formulation, not only the molecule

The representations above describe a single molecule, yet a drug product is a mixture of a drug and one or more excipients, processed under defined conditions. Representing that whole system is the central and still unsolved problem of formulation modelling. In practice, studies concatenate descriptors or fingerprints of the drug with descriptors of the excipients, encode categorical excipient identity as indicator variables, and append process parameters such as temperature, speed, and ratio as additional numerical features. This concatenation is pragmatic but crude: it does not capture the physical interaction between drug and polymer, it treats mixture composition as a set of independent inputs rather than a constrained blend, and it forces every study to invent its own encoding, which is one reason models rarely transfer between laboratories. The tablet, spray drying, and suspension studies discussed above show that formulation and process features frequently carry more predictive weight than the molecular descriptors of the drug, which underscores that a molecule-only representation is insufficient for the product[17,24,25].

4.5 Feature selection and interpretability

When hundreds or thousands of features accompany a few hundred data points, the number of variables can approach or exceed the number of observations, and indiscriminate use invites chance correlation. Feature selection, whether by filter, wrapper, or embedded methods, reduces the space to the informative variables and is a standard and necessary step.[34] Selection must be performed within the cross-validation loop rather than on the full dataset, because selecting features using the test data leaks information and inflates apparent performance, a subtle error that remains common. Interpretability has become a parallel requirement, driven both by scientific need and by regulatory expectation. Model-agnostic attribution methods, of which the Shapley additive explanation framework is the most used, assign each feature a contribution to a prediction and allow a model to be interrogated rather than trusted blindly.[35] Feature attribution is what allowed the suspension study to name stabiliser

concentration as the controlling variable, converting a predictive model into mechanistic insight. A comparison of the three representation families and their trade-offs is given in Table 2.

Table 2. Molecular representation families used as model inputs in formulation property prediction.

Representation	Examples and tools	Strengths	Limitations	Typical use
Expert descriptors	Constitutional, topological, electronic, geometric; Mordred [27], PaDEL [28]	Interpretable and mechanistically meaningful; work with small datasets	Informative set unknown in advance; require selection	Small formulation datasets; regulatory contexts
Structural fingerprints	MACCS keys [29], extended-connectivity fingerprints [30]	Fast; no descriptor selection; capture local substructure	Sparse; less interpretable; subject to bit collisions	Similarity and stability classification
Learned representations	Message-passing graph networks [31,32], SMILES transformers [33]	Optimised for the target property; strong when data are abundant	Data-hungry, opaque, and limited in transfer	Large libraries such as lipid nanoparticle screens

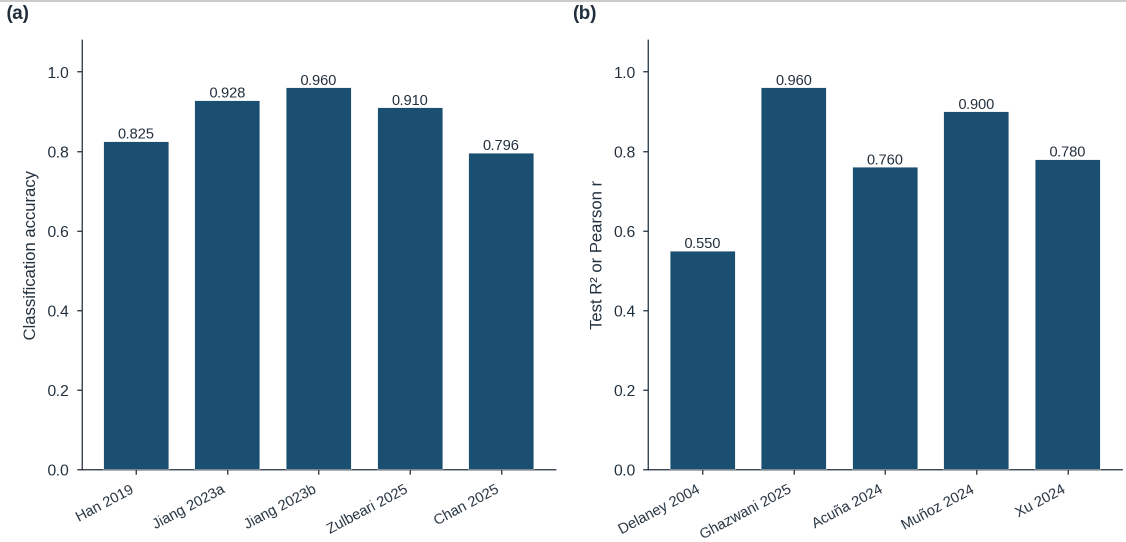

Figure 2. Reported predictive performance of representative machine learning models for formulation property prediction, compiled from the primary studies in Table 1. (a) Classification accuracy for stability and hit-identification tasks: amorphous solid dispersion physical stability (Han 2019 [14]), hot-melt-extrusion amorphization (Jiang 2023a [15]) and chemical stability (Jiang 2023b [15]), nanosuspension stability (Zulbeari 2025 [25]), and lipid nanoparticle top-performer classification (Chan 2025 [20]). (b) Coefficient of determination on the test set (R^2) or Pearson correlation coefficient (r) as reported for regression tasks: aqueous solubility (Delaney 2004, test R^2 [10]), tablet disintegration time (Ghazwani 2025 [18]), Caco-2 permeability (Acuña 2024 [23]), spray-dried particle size (Muñoz 2024, Formulation A [24]), and lipid nanoparticle potency (Xu 2024, Pearson r [19]). Values are the headline figures reported by each study under its own dataset and validation split and are not the product of a common benchmark, so cross-study comparison is indicative only.

5. Validation: what a trustworthy model must demonstrate

A model that predicts a formulation property will be used to make or defer an experiment, and in a regulated setting to support a filing, so the question is not whether it fits its training data but whether its predictions can be trusted on new inputs. The methodology for answering that question was developed for classical structure-property models and applies directly to their machine learning successors.

5.1 The established principles

The Organisation for Economic Co-operation and Development codified five principles for a regulatory-acceptable structure-property model: a defined endpoint, an unambiguous algorithm, a defined applicability domain, appropriate measures of goodness-of-fit, robustness and predictivity, and, where possible, a mechanistic interpretation.[36] These principles are the reference standard against which the reliability of a formulation model should be judged, and they map cleanly onto the recurring weaknesses in the current literature: endpoints defined inconsistently across studies, algorithms and preprocessing reported incompletely, applicability domains rarely stated, and predictivity measured internally rather than externally.

5.2 Internal fitting is not external predictivity

The single most important lesson of the classical literature is that strong internal statistics do not guarantee predictive power. A high leave-one-out cross-validation coefficient can coexist with poor performance on genuinely external compounds, a point demonstrated forcefully in the caution against reliance on that internal metric.[37] The recommended discipline is unambiguous: validate on an external test set held out from all model development, subject the model to response randomisation to confirm the correlation is not chance, and report the external predictivity explicitly.[38,39] Much of the formulation literature reports cross-validated performance on a single dataset without a true external set, which is why the headline accuracies of Section 3 should be read as upper bounds rather than as guarantees of field performance.

5.3 How the data are split

The way a dataset is partitioned determines what a validation measures. A random split of a dataset containing close analogues places near-identical compounds in both training and test sets, so the model is effectively tested on what it has already seen and its reported accuracy is optimistic. Splitting by molecular scaffold, using the framework definition of Bemis and Murcko, forces the test set to contain structurally distinct compounds and yields a harder, more realistic estimate, and standardised benchmarks now recommend it.[40,41] The dissimilarity-based split used in the solid dispersion stability study is an example of the same principle applied deliberately.[14] Related hazards include activity cliffs, pairs of very similar structures with very different properties that break the assumption that similar molecules behave similarly and that degrade prediction locally,[42] and temporal or

data-leakage effects when curation or feature selection touches the test data. In formulation datasets assembled from many publications, the same drug frequently recurs across records, and unless splitting respects that structure the model can memorise the drug rather than learn the property.

5.4 Applicability domain and honest metrics

No model is universally valid, and the applicability domain defines the region of chemical and formulation space where predictions are supported by the training data. Methods to define it range from simple descriptor-range and bounding-box approaches to leverage-based measures read from a Williams plot and distance-to-nearest-neighbour criteria, and the choice materially changes which predictions are deemed reliable.[43] A prediction for an input outside the domain should be flagged as an extrapolation, not reported with false confidence, and the growth of prospective error on out-of-distribution compounds in the binary-solvent solubility model is exactly the behaviour the domain is meant to fence off.[13] The metrics reported alongside must also be chosen to reflect agreement rather than mere correlation. A correlation coefficient can be high while predictions are systematically biased, which is why stricter measures such as the concordance correlation coefficient, and inspection of the observed-versus-predicted scatter, are recommended for external evaluation.[44] Taken together, an external and structure-aware test, an explicitly stated applicability domain, and metrics that measure agreement rather than correlation define a validation that a reviewer can accept and a regulator can assess.

6. Discussion

Read together, the studies surveyed here show a field that has proven its concept many times over and now faces the harder task of becoming dependable. The recurring limitations are not failures of individual authors but structural features of how the work is done.

The first and deepest is data. Formulation datasets are small, typically hundreds of records, and they are assembled from heterogeneous sources with inconsistent measurement conditions and reporting. Reported accuracies near the ceiling on such data are fragile, because the same drug recurs across records and because internal validation on a single dataset flatters the model. The most convincing studies in this review are those that either closed the loop with prospective experiments, as the lipid nanoparticle platforms did, or that deliberately hardened their validation with dissimilarity-based splitting.[14,19,20] The field would benefit more from shared, standardised, openly reported datasets than from any new algorithm, a point the regulatory guidance implicitly reinforces by demanding that model credibility be established for a defined context of use.[8,45]

The second is representation. A drug product is a processed mixture, and the field still lacks a principled way to represent it. The prevailing practice of concatenating molecular features with one-hot excipient labels and loose process

parameters is expedient but limits transfer, because every laboratory encodes its formulations differently and because the encoding ignores the physics of drug-excipient interaction. That formulation and process features often outweigh molecular descriptors in importance is a signal that the molecule-centric inheritance from drug discovery is a poor fit for the product.

The third is validation and reproducibility. The principles for trustworthy structure-property models are decades old and well specified, yet applicability domains are seldom stated, external test sets are often absent, and feature selection is not always kept clear of the test data.[36-38] Interpretability, increasingly expected by both scientists and regulators, is unevenly reported. These are not exotic requirements, and their inconsistent application is the clearest gap between the literature and decision-grade practice. The convergence of these models with quality-by-design frameworks, in which a defined design space and control strategy are already the language of development, offers a natural home for applicability-domain thinking and for the documentation that regulatory credibility assessment will require.[45,46]

A balanced reading also resists overcorrection. The pull toward ever more complex learned representations is not always warranted: on the small, noisy datasets typical of formulation, interpretable descriptor models are frequently as accurate and far more defensible than graph or transformer models, and the trade-off should be chosen deliberately rather than by fashion.[12] Complexity that cannot be validated on the available data is a liability, not an advance.

7. Future directions

Three concrete developments would move the field forward. First, the community needs shared formulation datasets with standardised descriptions of composition and process, curated to the reliability standard already applied to solubility data, so that models can be trained and benchmarked on common ground rather than on private collections.[11,41] Second, representation research should target the formulation as a system, developing encodings for mixtures and processes and for drug-excipient interaction that transfer across laboratories, rather than refining molecule-only representations imported from discovery. Third, reporting should be standardised around the established validation principles: every formulation model should state its applicability domain, report external predictivity with a metric that reflects agreement, keep feature selection inside the validation loop, and provide feature-level interpretation, so that a reader can judge where the model can be trusted.[36,43,44] Prospective experimental validation, still rare, should become the expected final step, as it already is in the strongest nanoparticle studies.[19,20] Alignment with quality-by-design and with the emerging regulatory credibility frameworks would give these practices an institutional anchor.[8,46]

8. Conclusion

Artificial intelligence can now predict a wide range of drug product properties, from solubility

and dissolution to amorphous stability, tablet mechanics, and nanoparticle performance, and in the best cases the predictions have guided successful experiments. The evidence assembled here supports a clear conclusion: the discipline is mature in its algorithms and immature in its data, its representation of the formulation, and its validation. The path to dependable, decision-grade, and regulator-ready models runs not through more powerful learning methods but through shared standardised data, representations that describe the product rather than the molecule alone, and the disciplined external, domain-aware, interpretable validation that the science of structure-property modelling has demanded for decades. A formulation model earns trust the same way a formulation does, by performing reliably outside the conditions in which it was made.

Declarations

Conflicts of interest: None declared.

References

1. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Adv Drug Deliv Rev. 2021;175:113806. doi:10.1016/j.addr.2021.05.016. PMID 34019959.
2. Hansch C, Fujita T. rho-sigma-pi Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc. 1964;86(8):1616-1626. doi:10.1021/ja01062a035.
3. Free SM Jr, Wilson JW. A Mathematical Contribution to Structure-Activity Studies. J Med Chem. 1964;7(4):395-399. doi:10.1021/jm00334a001.
4. Boczar D, Michalska K. A Review of Machine Learning and QSAR/QSPR Predictions for Complexes of Organic Molecules with Cyclodextrins. Molecules. 2024;29(13):3159. doi:10.3390/molecules29133159.
5. Suriyaamporn P, Pamornpathomkul B, Patrojanasophon P, Ngawhirunpat T, Rojanarata T, Opanasopit P. The Artificial Intelligence-Powered New Era in Pharmaceutical Research and Development: A Review. AAPS PharmSciTech. 2024;25(6):188. doi:10.1208/s12249-024-02901-y. PMID 39147952.
6. Warke S, Katari O, Jain S. Current Status on the Convergence of Artificial Intelligence and Formulation Development in Industry: A Review. AAPS PharmSciTech. 2025;27(1):44. doi:10.1208/s12249-025-03296-0.
7. US Food and Drug Administration. Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products; Availability. Fed Regist. 2023 May 11;88(91):30313. Docket FDA-2023-N-0743.
8. US Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological

Products; Draft Guidance for Industry. Fed Regist. 2025 Jan 7;90(4):1157. Docket FDA-2024-D-4689.

9. European Medicines Agency. Reflection paper on the use of artificial intelligence in the lifecycle of medicines. EMA/CHMP/CVMP/83833/2023. Amsterdam: EMA; 2024.
10. Delaney JS. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. J Chem Inf Comput Sci. 2004;44(3):1000-1005. doi:10.1021/ci034243x.
11. Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Sci Data. 2019;6(1):143. doi:10.1038/s41597-019-0151-1.
12. Zheng T, Mitchell JBO, Dobson S. Revisiting the Application of Machine Learning Approaches in Predicting Aqueous Solubility. ACS Omega. 2024;9(32):35209-35222. doi:10.1021/acsomega.4c06163. PMID 39157153.
13. Bao Z, Tom G, Cheng A, Watchorn J, Aspuru-Guzik A, Allen C. Towards the prediction of drug solubility in binary solvent mixtures at various temperatures using machine learning. J Cheminform. 2024;16(1):117. doi:10.1186/s13321-024-00911-3.
14. Han R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, et al. Predicting physical stability of solid dispersions by machine learning techniques. J Control Release. 2019;311-312:16-25. doi:10.1016/j.jconrel.2019.08.030. PMID 31465824.
15. Jiang J, Lu A, Ma X, Ouyang D, Williams RO 3rd. The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion. Int J Pharm X. 2023;5:100164. doi:10.1016/j.ijpx.2023.100164. PMID 36798832.
16. Lee H, Kim J, Kim S, Yoo J, Choi GJ, Jeong YS. Deep Learning-Based Prediction of Physical Stability considering Class Imbalance for Amorphous Solid Dispersions. J Chem. 2022;2022:4148443. doi:10.1155/2022/4148443.
17. Akseli Ilgaz, Xie J, Schultz L, Ladyzhynsky N, Bramante T, He X, et al. A Practical Framework Toward Prediction of Breaking Force and Disintegration of Tablet Formulations Using Machine Learning Tools. J Pharm Sci. 2017;106(1):234-247. doi:10.1016/j.xphs.2016.08.026.
18. Ghazwani M, Hani U. Prediction of tablet disintegration time based on formulations properties via artificial intelligence by comparing machine learning models and validation. Sci Rep. 2025;15(1):13789. doi:10.1038/s41598-025-98783-6.
19. Xu Y, Ma S, Cui H, Chen J, Xu S, Gong F, et al. AGILE platform: a deep learning powered approach to accelerate LNP development for mRNA delivery. Nat

Commun. 2024;15(1):6305. doi:10.1038/s41467-024-50619-z.

20. Chan A, Kirtane AR, Qu QR, Yang YS, Yang KS, Maiorino L, et al. Designing lipid nanoparticles using a transformer-based neural network. Nat Nanotechnol. 2025;20:1491-1501. doi:10.1038/s41565-025-01975-4.

21. Hoseini B, Jaafari MR, Golabpour A, Momtazi-Borojeni AA, Karimi M, Eslami S. Application of ensemble machine learning approach to assess the factors affecting size and polydispersity index of liposomal nanoparticles. Sci Rep. 2023;13(1):18012. doi:10.1038/s41598-023-43689-4.

22. He Y, Ye Z, Liu X, Wei Z, Qiu F, Li HF, et al. Can machine learning predict drug nanocrystals? J Control Release. 2020;322:274-285. doi:10.1016/j.jconrel.2020.03.043. PMID 32234511.

23. Acuña-Guzman V, Montoya-Alfaro ME, Negrón-Ballarte LP, Solis-Calero C. A Machine Learning Approach for Predicting Caco-2 Cell Permeability in Natural Products from the Biodiversity in Peru. Pharmaceuticals (Basel). 2024;17(6):750. doi:10.3390/ph17060750.

24. Muñoz López CA, Peeters K, Van Impe JFM. Data-Driven Modeling of the Spray Drying Process. Process Monitoring and Prediction of the Particle Size in a Pharmaceutical Production Process. ACS Omega. 2024;9(9):9741-9754. doi:10.1021/acsomega.3c08032.

25. Zulbeari N, Wang F, Mustafova SS, Parhizkar M, Holm R. Machine learning strengthened formulation design of pharmaceutical suspensions. Int J Pharm. 2025;668:124967. doi:10.1016/j.ijpharm.2024.124967.

26. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Weinheim: Wiley-VCH; 2009. doi:10.1002/9783527628766.

27. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4. doi:10.1186/s13321-018-0258-y. PMID 29411163.

28. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-1474. doi:10.1002/jcc.21707. PMID 21425294.

29. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42(6):1273-1280. doi:10.1021/ci010132r. PMID 12444722.

30. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742-754. doi:10.1021/ci100050t. PMID 20426451.

31. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for

Quantum Chemistry. Proc 34th Int Conf Mach Learn (ICML), PMLR. 2017;70:1263-1272.

32. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model. 2019;59(8):3370-3388. doi:10.1021/acs.jcim.9b00237. PMID 31361484.

33. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv:2010.09885. 2020.

34. Guyon Isabelle, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003;3:1157-1182.

35. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst. 2017;30:4765-4774.

36. Organisation for Economic Co-operation and Development. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2007. doi:10.1787/9789264085442-en.

37. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269-276. doi:10.1016/S1093-3263(01)00123-1. PMID 11858635.

38. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010;29(6-7):476-488. doi:10.1002/minf.201000061.

39. Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007;26(5):694-701. doi:10.1002/qsar.200610151.

40. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem. 1996;39(15):2887-2893. doi:10.1021/jm9602928. PMID 8709122.

41. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-530. doi:10.1039/c7sc02664a. PMID 29629118.

42. Stumpfe D, Bajorath J. Evolving Concept of Activity Cliffs. ACS Omega. 2019;4(11):14360-14368. doi:10.1021/acsomega.9b02221. PMID 31528788.

43. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules. 2012;17(5):4791-4810. doi:10.3390/molecules17054791. PMID 22534664.

44. Chirico N, Gramatica P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. J Chem Inf Model. 2011;51(9):2320-2335. doi:10.1021/ci200211n.

45. Panwar R, Mishra A, Sahu A, Quadri SN, Abdin MZ, Fatima S, et al. Implementing QbD for Nano-Pharmaceuticals and Complex Formulations to Achieve Predictable and High-Quality Outcomes. AAPS PharmSciTech. 2025;27(1):71. doi:10.1208/s12249-025-03308-z.

46. Zhu Z. Intelligent information management enables quality-by-design in pharmaceutical production. Sci Rep. 2025;15(1):44201. doi:10.1038/s41598-025-27879-w.