Journal of Advanced Pharmaceutical Sciences and Natural Products

Journal of Advanced Pharmaceutical Sciences and Natural Products (JAPSNP)

MACHINE LEARNING GUIDED FORMULATION DESIGN: FROM DESCRIPTOR-BASED PREDICTION OF SOLUBILITY AND STABILITY TO AUTONOMOUS SELF-DRIVING FORMULATION LABORATORIES

Rajat Bhattacharjee*

Assistant Professor, Department of Pharmaceutics, Bengal College of Pharmaceutical Technology, Dubrajpur, Birbhum, 731123

Corresponding author:
Rajat Bhattacharjee

Email: rajatbhattacharjee23@gmail.com

ABSTRACT

Formulation development still consumes scarce drug substance, months of stability storage and a great deal of expert intuition. Machine learning (ML) offers a different economy, one in which experiments are chosen for the information they return rather than for their place in a factorial grid. This review traces that shift across three stages of maturity. The first is descriptor-based prediction, where classical equations such as the General Solubility Equation and ESOL gave way to fingerprint, graph and physics-informed models trained on curated sets such as AqSolDB. The evidence examined here shows that aqueous solubility models have reached a floor set largely by the reproducibility of the measurements themselves, which ranges from about 0.15 log unit for standardised potentiometric methods to roughly 1 log unit across heterogeneous literature. The second stage is stability prediction, covering amorphous solid dispersions, glass-forming ability, co-crystal formation, excipient compatibility and protein formulation, where classifiers reach 80 to 97% accuracy on in vitro endpoints but rarely report applicability domains. The third stage closes the loop: Bayesian optimisation and active learning coupled to robotic platforms, which have located high-solubility curcumin formulations after testing about 3% of a 7,776-member design space and produced in-specification tablets within six hours of material characterisation. Only in vitro and in silico evidence is considered; studies relying on human or animal data were excluded by design. The review closes with a critical appraisal of data quality, validation practice and regulatory credibility, drawing a lesson from the correction of a prominent autonomous-synthesis claim, and sets out what is needed before autonomous formulation laboratories can be trusted with regulated products.

Keywords: machine learning; drug formulation; solubility; drug stability; amorphous solid dispersion; Bayesian optimisation; self-driving laboratory; quality by design.

1. Introduction

A new chemical entity reaches the formulation scientist with three unwelcome properties more often than not. It is poorly soluble, it is available in gram quantities at best, and its long-term stability is unknown. The classical response, codified in the quality-by-design framework of ICH Q8(R2), is to map a design space through structured experimentation and to confirm stability under the storage conditions of ICH Q1A(R2) [1,2]. The framework is sound. Its cost is the problem. A factorial design over five excipients at three levels already demands hundreds of runs, and real-time stability data arrive on a calendar measured in months.

Machine learning changes the arithmetic in two distinct ways, and the distinction organises this review. In the first, a model trained on historical data predicts a property, such as solubility or physical stability, before any material is spent. In the second, a model decides which experiment to run next, so that the data needed to answer a question shrink as the model learns. The first is prediction; the second is decision. Self-driving laboratories (SDLs) are the point where decision-making is handed to an algorithm that also controls the robot [3,4].

Figure 1 sets out the three stages around which this review is organised and shows how they close into a single loop.

Several reviews have mapped parts of this territory, from computational pharmaceutics as a discipline [5] to ML-directed formulation [6] and the broader SDL literature in chemistry and materials science [7]. What is missing is an account that follows a single thread from the molecular descriptor to the autonomous experiment and asks, at each step, how good the evidence actually is. That is the purpose here. The argument advanced is that the limiting factor in ML-guided formulation has moved. It is no longer the learning algorithm. It is the quality, scope and reporting of the data on which the algorithms are trained and against which they are judged, and closed-loop experimentation matters chiefly because it is the most efficient way yet devised to generate such data.

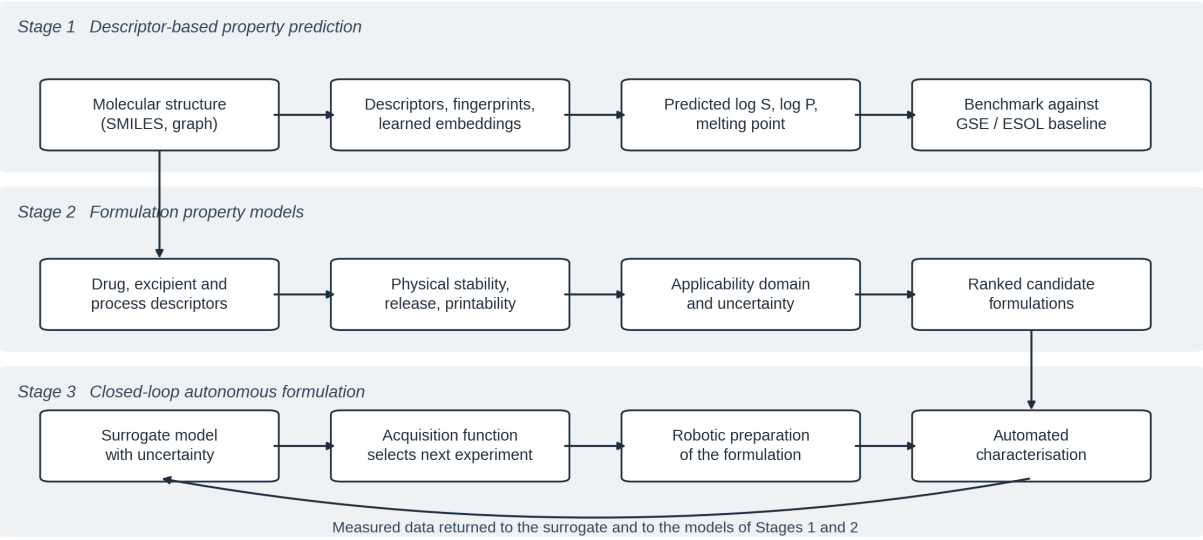

Figure 1. The three stages of machine learning guided formulation design form a single closed loop rather than a linear pipeline. Stage 1 predicts molecular properties, such as aqueous solubility and melting point, from structure-derived descriptors, fingerprints or learned graph embeddings, and is judged against simple physical baselines. Stage 2 extends the same machinery to mixtures by concatenating descriptors of drug, excipient and process, and returns ranked candidate formulations together with an applicability domain. Stage 3 replaces prediction with sequential decision making: a surrogate model carrying its own uncertainty proposes the next experiment through an acquisition function, a robotic platform prepares and characterises it, and the measured result is returned to the surrogate and to the models of Stages 1 and 2. Arrows indicate the direction of information flow. The diagram is original to this review and summarises the workflow described in Sections 3 to 7.

2. Scope and approach

This is a narrative review. Peer-reviewed literature, benchmark datasets and regulatory documents published between 2001 and 2026 were identified through PubMed, Europe PMC, publisher databases and snowballing from the reference lists of key reviews, using combinations of the terms *machine learning*, *solubility*, *stability*, *solid dispersion*, *formulation*, *Bayesian optimisation* and *self-driving laboratory*. Every reference was checked against its publisher or registry record.

One boundary was applied throughout. Studies whose findings rest on human or animal data, including pharmacokinetic, bioavailability and in vivo efficacy outcomes, were excluded. The evidence considered is therefore computational, physicochemical or in vitro. This choice is deliberate rather than merely restrictive, since in vitro endpoints such as equilibrium solubility, amorphous stability and release rate are exactly the quantities that ML-guided formulation optimises in the laboratory.

No meta-analysis was attempted. The included studies report different endpoints (log S, classification accuracy, mean absolute error on release fraction, number of experiments to an optimum) on datasets that are neither independent nor comparable in their splitting protocols. A pooled estimate across them would answer no sensible question, and the synthesis is therefore narrative, supported by two structured tables.

3. Representing a molecule and a formulation

Every predictive model begins with a decision about what the computer is allowed to see. The earliest solubility models saw very little, and did remarkably well. The General Solubility Equation (GSE) of Jain and Yalkowsky estimates the aqueous solubility of a nonelectrolyte from only its melting point and octanol-water partition coefficient, and it was validated across 580 organic compounds of pharmaceutical, environmental and industrial relevance [8]. Delaney's ESOL removed the need for a measured melting point by regressing 2,874 solubility values on nine calculated properties, of which log P, molecular weight, aromatic proportion and rotatable bond count carried the weight [9]. Both are interpretable, and both remain baselines against which more elaborate models ought to be compared, and often are not.

Three families of representation followed. Hand-crafted descriptors extend the ESOL idea to thousands of computed properties. Circular fingerprints such as the extended-connectivity fingerprint encode substructures around each atom without a predefined dictionary [10]. Learned representations, most prominently directed message-passing neural networks, build a molecular vector from the graph itself; on 19 public and 16 proprietary datasets such models matched or outperformed fixed descriptors [11]. The MoleculeNet benchmark offered a useful caution against treating architecture as destiny, reporting that for physical and biophysical endpoints the choice of featurisation can matter more than the choice of learning algorithm [12].

A formulation is harder to represent than a molecule. It is a mixture, often of a drug with a polymer, surfactant or co-former, defined also by composition and process history. Most formulation models therefore concatenate descriptors of each component with weight fractions and process variables, an approach used for solid dispersions [13] and for printed dosage forms [14]. It works, but it discards the physics of interaction between components, which is why formulation datasets must be far larger than single-molecule datasets to generalise comparably.

4. Solubility: a prediction problem approaching its ceiling

Solubility is the most intensively modelled property in pharmaceutics and also the most instructive, because its history shows a field discovering the limits of its own data.

4.1 From curated sets to challenges

The available training data have grown and, more importantly, have been curated. AqSolDB merged nine public sources into 9,982 unique compounds with standardised units and a reliability grade [15]. The OPERA suite, built on curated PHYSPROP data under OECD validation principles, delivered physicochemical models with test-set R^2 between 0.71 and 0.96 [16,17]. For organic solvents, which matter for crystallisation and for lipid-based and spray-dried formulations, SolProp assembled more than 5,000 experimental values and supported predictions at temperatures up to 550 K and beyond [18].

Against this improving supply of data, the community organised two blind challenges. The first asked participants to predict 32 molecules from a training set of 100 intrinsic solubilities measured potentiometrically at 25 °C and 0.15 M ionic strength, with measurements rejected above a 4% coefficient of variation [19]. The second, more revealing, split its test compounds into a tight set and a contentious set according to how well independent laboratories agreed: reproducibility averaged about 0.17 log unit for the tight set and about 0.62 log unit for the contentious one [20]. The second challenge thus measured two things at once, the predictions and the yardstick.

4.2 The reproducibility ceiling

The uncomfortable finding of the last decade is that solubility models now compete with the noise in their own reference values. Palmer and Mitchell put the question directly and answered it against the prevailing assumption: literature solubility data carry roughly 0.6 to 0.7 log unit of uncertainty against less than 0.05 for carefully standardised potentiometric measurements, yet models trained on the cleaner data were not correspondingly more accurate, which points to the modelling approach rather than data noise as the binding constraint [21]. Interlaboratory comparison sharpens the picture from the measurement side: average between-laboratory standard deviations of 0.15 log unit for potentiometric and 0.24 log unit for shake-flask determinations, with the two methods agreeing at an RMSE of 0.34 log unit [22]. A more recent analysis of aggregated public data found interlaboratory spread reaching 1.0 log unit for some compounds and argued that this sets a lower bound on achievable prediction accuracy; external validation of models trained on such data gave RMSEs between 1.74 and 2.17 log units [23].

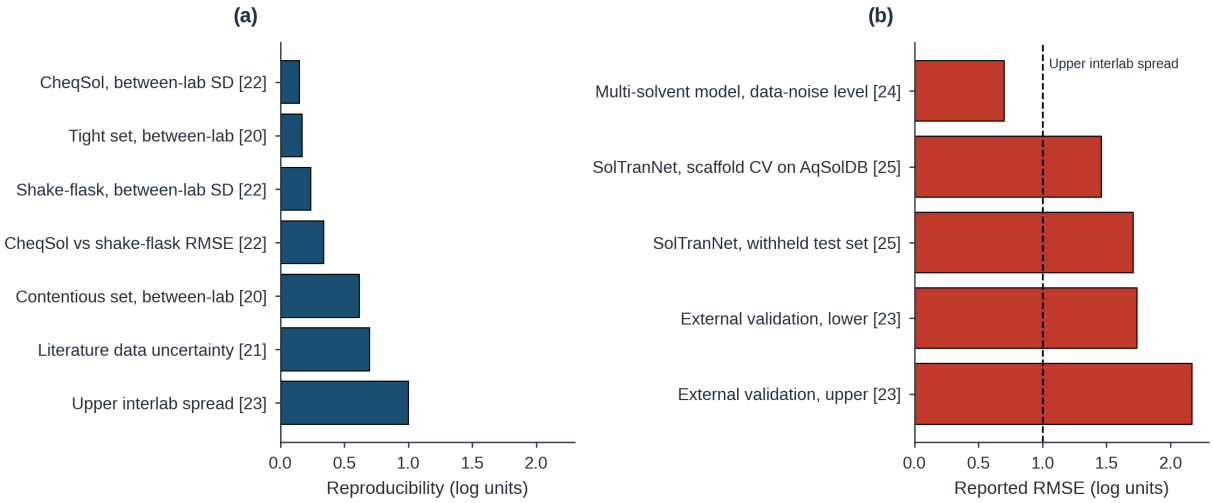

Figure 2. Reported errors of aqueous solubility models are large relative to, and bounded below by, the reproducibility of the reference measurements themselves. (a) Reproducibility of experimental solubility determinations, expressed in log units: between-laboratory standard deviations for potentiometric (CheqSol) and shake-flask methods and the root mean square error between the two methods [22], between-laboratory agreement for the tight and contentious compound sets of the second solubility challenge [20], the uncertainty attributed to heterogeneous literature data [21], and the upper interlaboratory spread reported for aggregated public data [23]. (b) Reported model errors on the same axis: the data-noise level approached by a physicochemically informed multi-solvent model [24], scaffold-split cross-validation and withheld-test errors for a transformer model [25], and the range of external validation errors reported for models trained on aggregated public data [23]. The dashed line marks the upper interlaboratory spread from panel (a). All values are as reported in the cited sources; no error bars are shown because these are single reported summary statistics rather than distributions measured here.

4.3 What a defensible solubility model reports

Three reporting practices separate models that transfer from models that do not, and none of them is novel. Splitting must be by scaffold or by cluster, since random splits on a chemically redundant dataset measure memorisation. An

applicability domain must be declared, so that a prediction for a molecule outside the training chemistry is flagged rather than silently returned; this has been a formal expectation of validated structure-property models since the OECD principles were issued [17]. And the comparison baseline must include a simple model. Where a gradient-boosted ensemble on two thousand descriptors cannot beat the General Solubility Equation on the same split, that fact is a result and belongs in the paper.

Interpretability deserves more scepticism than it receives. Shapley additive explanations give a uniquely determined attribution of a prediction across input features [27] and are now routine for solubility and stability models. What they explain is the model, not the physical system. A high attribution to topological polar surface area states what the learned function uses; it becomes a mechanistic claim only when independent evidence supports it.

5. Stability: the harder and more valuable problem

Solubility can be measured in a day. Stability cannot, and that asymmetry makes it the more valuable target for prediction: a model that correctly rejects an unstable amorphous dispersion saves six months, not six hours.

5.1 Physical stability of amorphous systems

Amorphous solid dispersions raise apparent solubility at the cost of thermodynamic stability, so their design is a problem of predicting recrystallisation. Two lines of work converged on this. The experimental line classified 51 organic molecules by their behaviour on melt-quenching into three crystallisation classes and linked poor glass-forming ability to low molecular weight and structural rigidity [28]. The computational line took the descriptor route: support vector machine models trained on 131 melt-quenched drugs recovered the classification with 86% accuracy on the crystallisation dataset and confirmed a molecular weight threshold near 300 for reliable glass formation [29].

Formulated dispersions add the polymer. A random forest trained on 646 stability observations described by more than twenty molecular descriptors reached 82.5% classification accuracy and was checked experimentally against 17β-estradiol dispersions in polyvinylpyrrolidone [13]. The same group extended the approach into a publicly usable platform predicting physical stability, dissolution type and dissolution rate and screening drug-polymer pairs in silico [30]. More recent work combining high-throughput screening with machine learning across 1,272 binary and ternary systems, of which 188 were amorphous, reported 96.7% classification accuracy for whether an amorphous dispersion forms at all [31]. That headline figure needs its class imbalance read alongside it, since the crystalline class dominates by roughly six to one, which is precisely why the accompanying precision of about 88% and F1 of 84% are the more informative numbers.

5.2 Crystallisation and co-crystals

Whether a molecule will crystallise well is a question of manufacturability, and it has been modelled directly from structure with classifiers

that separate good crystal formers from poor ones and predict how a synthetic modification shifts that behaviour [32]. Co-crystal formation, a standard route to improved dissolution, has been predicted by neural networks trained on the co-crystal content of the Cambridge Structural Database, achieving roughly 80% accuracy for untested co-former pairs [33]. Both models share a structural weakness worth naming: they are trained on the crystallographic record, which records successes far more completely than failures. A model trained on what has been deposited learns what crystallographers chose to deposit.

5.3 Chemical stability and excipient compatibility

Chemical degradation has been addressed for two decades by isoconversion methods rather than by learning. The accelerated stability assessment programme, built on a humidity-corrected Arrhenius relationship, estimates shelf life from protocols lasting about two weeks [34,35]. These remain the practical standard, and any ML approach to chemical stability must be judged against them rather than against no method at all.

Excipient compatibility is where learning has recently added something. A stacked model combining mol2vec embeddings with two-dimensional descriptors reported accuracy of 0.98, an area under the curve of 0.93 and a Matthews correlation coefficient of 0.86, and in experimental checking identified 10 of 12 known incompatibilities against 3 of 12 for an earlier rule-based tool [36]. The Matthews coefficient is the number to trust here, since accuracy on an imbalanced compatibility dataset is close to uninformative.

5.4 Biologics formulation

Protein formulation is a search over buffer, pH, ionic strength and excipients for a composition that keeps the molecule monomeric for two years. Neural networks trained on accelerated-stress and high-throughput biophysical measurements have been used to predict real-time monomer retention, validated by leaving each protein out of training in turn, which is the appropriate test when the question is whether the model transfers to a new molecule [37]. The decision-making variant is more striking. Bayesian optimisation identified the most thermally stable formulation for three tandem single-chain variable fragment variants in 25 experiments, under a third of the runs a classical design of experiments approach required [38]. That result belongs as much to section 7 as to this one, and it is the clearest single demonstration in the formulation literature that choosing experiments beats predicting outcomes when material is scarce [39].

5.5 Process and printability

Manufacturability has been modelled in the same descriptor-based idiom. Models trained on 614 drug-loaded formulations built from 145 excipients predicted three-dimensional printability with 76% accuracy, and extrusion and printing temperatures to within 8.9 °C and 8.3 °C [14]. A larger study across 968 formulations drawn from 114 publications reached up to 93% accuracy for filament extrusion and predicted

drug release time to within about 24 minutes [40]. Release-rate prediction has also been demonstrated for long-acting injectables: a gradient-boosted model trained on 181 release profiles comprising 3,783 measurements across 43 drug-polymer pairs achieved a test mean absolute error of 0.114 in released fraction and guided the design of two new poly(lactic-co-glycolic acid) formulations, evaluated in vitro [41].

Table 1. Representative approaches to aqueous and organic-solvent solubility prediction, with the datasets and reported performance on which each rests. All entries are computational or in vitro; none involves human or animal data. GSE, General Solubility Equation; SD, standard deviation; CV, cross-validation; MAE, mean absolute error.

Approach	Data	Validation reported	Reported performance	Ref.
GSE, two physical inputs	580 nonelectrolytes	External set	Interpretable physical baseline	8
ESOL, 9 calculated properties	2,874 measured values	Held-out test	Comparable to GSE without melting point	9
Curated reference database	9,982 unique compounds	Reliability grading	Standardised benchmark resource	15
OECD-compliant descriptor models	Curated PHYSPROP	5-fold CV and external	Test R^2 0.71 to 0.96 (mean 0.82)	16
Quality-weighted descriptor selection	AqSolDB subsets	Leave-one-out	MAE 0.348, RMSE 0.483, R^2 0.94	26
Physicochemically informed ML, multi-solvent	Mixed aqueous and organic	Held-out test	Approaches data-noise level (~0.7 log)	24
Transformer, aqueous	AqSolDB	Scaffold-split CV	RMSE 1.459 CV, 1.711 withheld	25
Solvent and temperature-dependent model	>5,000 values (SolProp)	External	Valid to 550 K and above	18
Aggregated-data reassessment	Public aqueous data	External validation	RMSE 1.74 to 2.17 log units	23

Table 2. Machine learning applied to stability, compatibility and manufacturability endpoints. Every study listed is in vitro or in silico. ASD, amorphous solid dispersion; MCC, Matthews correlation coefficient; RF, random forest; SVM, support vector machine; BO, Bayesian optimisation.

Endpoint	Data	Model	Reported performance	Ref.
Crystallisation tendency from the melt	51 molecules	Experimental classification	Three-class scheme, rigidity and low mass predict poor glass formation	28
Glass-forming ability	131 drugs	SVM on descriptors	86% of crystallisation dataset predicted	29
ASD physical stability	646 observations	RF among 8 algorithms	82.5% accuracy, confirmed experimentally	13
ASD formation, binary and ternary	1,272 systems (188 amorphous)	RF	96.7% accuracy, precision ~88%, F1 84%	31
Co-crystal formation	Cambridge Structural Database	Neural network	~80% accuracy on untested pairs	33
Drug-excipient compatibility	Curated compatibility set	Stacked mol2vec model	Accuracy 0.98, AUC 0.93, MCC 0.86	36
Protein monomer retention	Accelerated-stress biophysical data	Neural network	Leave-one-protein-out validation	37
Protein formulation optimisation	3 scFv variants	BO	Optimum in 25 experiments, under one third of design-of-experiments	38
3D printability	614 formulations, 145 excipients	Ensemble	76% accuracy; temperatures within 8.9 °C and 8.3 °C	14
3D printing performance	968 formulations	Neural network	Up to 93% accuracy; release time ±24 min	40
Long-acting injectable release	181 profiles, 3,783 measurements	Gradient boosting	Test MAE 0.114 released fraction	41

6. From predicting a property to choosing an experiment

Everything above is prediction. A model is fitted to data that already exist and is asked what it thinks about a molecule it has not seen. Formulation development, however, is not a prediction task. It is a sequential decision task under a budget of material, time and instrument hours, and the appropriate machinery is different.

Bayesian optimisation is that machinery. A surrogate model, most often a Gaussian process, carries not only an estimate of the objective across the design space but an estimate of its own uncertainty; an acquisition function converts the two into a single proposal for the next experiment, balancing exploitation of promising regions against exploration of ignorant ones [42]. Chemistry has produced variants suited to its own awkwardness: optimisers that handle categorical variables such as solvent identity alongside continuous ones [43], and formulations of the problem that respect the constraints a laboratory actually imposes, such as excipients that cannot be combined or compositions that must sum to unity [44].

The benchmark that most changed opinion was a synthetic one. In a controlled comparison run as a game against expert chemists, Bayesian optimisation of a palladium-catalysed direct arylation reached better yields in fewer experiments and did so more consistently than human decision-making, before being applied to further reaction classes [45]. The result transfers to formulation with one adjustment: a formulation objective is usually multiple, since solubility, stability and manufacturability must be satisfied together, so multi-objective acquisition functions that return a Pareto set rather than a single optimum are the appropriate tool [46].

Active learning is the same idea with a different objective. Rather than optimising a property, the algorithm requests the experiments that would most reduce model uncertainty, which is the correct strategy when the goal is a model that transfers rather than a single good formulation. Applied on a robotic platform to polymer-protein hybrids, active learning identified copolymers that preserved or improved the activity of three chemically distinct enzymes under thermal stress [47]. The materials community formalised the closed-loop version of this at instrument scale, with on-the-fly Bayesian active learning driving a synchrotron beamline through cycles lasting seconds to minutes [48].

7. Self-driving formulation laboratories

An SDL is the point at which the decision algorithm is wired to the hardware that executes it, so that design, execution, analysis and redesign proceed without a human in the loop [4,7].

7.1 What the demonstrations established

The canonical demonstration is a mobile robot that worked autonomously for eight days, performing 688 experiments across a ten-dimensional space under batched Bayesian search and identifying photocatalyst mixtures about six times more active than the starting compositions [49]. A modular thin-film platform optimised hole mobility by varying composition

and processing together, establishing that process parameters belong inside the search space rather than outside it [50]. A community perspective drawn from academia, industry and national laboratories set out what such systems require in infrastructure, algorithms and skills [51].

Large language models have since been used as the planning layer, with one agent handling six task types including the optimisation of palladium-catalysed cross-couplings [52] and a related system linking a language model to eighteen expert chemistry tools [53]. These are best understood as an interface improvement of real consequence, lowering the barrier between a scientific question and a machine-executable protocol, rather than as a new source of chemical knowledge.

7.2 Closing the loop on formulation

Formulation-specific autonomy is younger and, for this review, more important. Three demonstrations mark the current state.

A semi-autonomous robotic formulator combining liquid handling with Bayesian optimisation searched a 7,776-member excipient design space by testing 256 combinations, about 3% of the space, and found seven curcumin formulations exceeding 10 mg/mL within days, an efficiency gain reported as roughly tenfold over exhaustive screening [54]. An automated crystallisation platform applied Bayesian optimisation to the cooling crystallisation of lamivudine, varying cooling rate, seed mass and supersaturation, and improved its objective by about 10% after a single iteration, which is notable less for the magnitude than for the demonstration that scale-up-relevant process variables can sit inside the loop [55]. Most consequential for solid dosage forms, a digital formulator coupled to a self-driving tableting line reduced the path from material characterisation to in-specification tablets to six hours while consuming about 65% less drug substance than conventional development [56].

The drug-substance saving is the figure that matters. Material scarcity, not instrument time, is the binding constraint in early formulation, and it is the constraint that closed-loop experimentation relieves most directly.

7.3 The cautionary case

Enthusiasm for autonomy should be tempered by the most instructive episode in the field. An autonomous inorganic synthesis laboratory reported producing a large number of predicted target compounds during 17 days of continuous operation [57]. Independent reanalysis argued that roughly two thirds of the claimed successes were better explained as known, compositionally disordered variants of the predicted ordered phases, attributing the error to automated Rietveld refinement applied without human judgement and to predictions that ignored disorder [58]. The paper was subsequently corrected, with its title amended, five compounds withdrawn from the count and the meaning of novelty clarified [57].

The lesson is not that autonomy failed. It is that autonomy removes the human from the analysis step as well as the decision step, and that

automated characterisation is the weakest link in the chain. In formulation the analogous risks are concrete: automated dissolution profiles misassigned through a baseline artefact, a powder diffraction pattern of an amorphous dispersion read as a halo when a weak Bragg reflection is present, or a turbidity reading taken as solubility. Any formulation SDL intended for regulated work needs orthogonal confirmation of its own analytics, and a defined point at which a human examines the raw signal.

8. Critical appraisal

Five weaknesses run across the literature reviewed here, and they are more alike than the diversity of endpoints suggests.

Data quality bounds everything. In solubility this is now quantified, with interlaboratory spread of 0.15 to 1.0 log unit depending on method and compound setting a floor beneath which reported accuracy should be disbelieved [22,23]. In stability the equivalent figure is unknown, because no comparable interlaboratory study of amorphous dispersion stability classification exists. That absence is itself a finding: stability models are being compared to each other without anyone knowing the reproducibility of the labels.

Negative results are missing. Formulation datasets are assembled from publications, and publications report formulations that worked. Co-crystal models trained on deposited structures, printability models trained on published prints and stability models trained on reported dispersions all inherit this asymmetry. Closed-loop platforms are valuable here for a reason rarely stated: they generate failures at the same fidelity as successes, and those failures are what supervised models lack.

Validation practice is uneven. Scaffold or cluster splitting, leave-one-molecule-out protocols and declared applicability domains appear in some of the strongest work reviewed here [37,25] and are absent elsewhere. Where a model reports only a random-split metric on a redundant dataset, its headline number is not comparable with anything.

Class imbalance is under-reported. Several stability classifiers report accuracies above 90% on datasets where one class predominates heavily [31]. Matthews correlation, balanced accuracy or the full confusion matrix should be mandatory for these endpoints, and precision on the minority class is usually the quantity a formulator cares about.

Interpretability is often overstated. Attribution methods explain a fitted function [27]. Treating a high-attribution descriptor as a mechanism, absent independent physical evidence, converts a modelling artefact into a scientific claim.

9. Regulatory and data-infrastructure outlook

Regulators have begun to set out how models used in drug development will be judged. A 2025 FDA draft guidance sets out a risk-based credibility assessment framework for artificial intelligence supporting regulatory decisions on drugs and biological products, scaling the evidence required to the consequence of the

model being wrong [59]. The European Medicines Agency's reflection paper, adopted in September 2024, takes a lifecycle view and stresses documentation, data governance and human oversight [60]. Neither is an obstacle to the work reviewed here. Both imply, however, that a model supporting a formulation decision must arrive with its training data, validation protocol, applicability domain and failure modes documented, a higher standard than most published formulation models meet.

Stability testing itself is under revision, with the consolidated ICH Q1 draft guideline endorsed at Step 2 in April 2025, merging the Q1A to Q1F series with Q5C [61]. How far modelled stability evidence can substitute for storage data remains the open question, and it is the question the field should be organising its evidence to answer.

Underneath all of this sits data infrastructure. The FAIR principles were written for exactly this situation, in which machines rather than people are the primary consumers of data [62]. A formulation dataset that is findable, accessible, interoperable and reusable, and that records failures and measurement uncertainty alongside successes, is worth more to the field than another architecture benchmarked on AqSolDB.

10. Future directions

Four specific things would move this field further than another round of model comparison.

First, an interlaboratory reproducibility study for a stability endpoint, on the model of the solubility challenges [20]. Ten laboratories, a shared set of drug-polymer dispersions, a common storage protocol and a common analytical decision rule would establish the label noise that every published stability classifier is implicitly claiming to beat.

Second, a public formulation dataset that includes failures, with composition, process parameters, raw analytical signals and uncertainty, deposited under FAIR terms. The closed-loop platforms now operating are its natural producers [54,56].

Third, reporting standards for formulation machine learning: mandatory scaffold or cluster splits, declared applicability domains, minority-class metrics, comparison against a simple physical baseline such as the General Solubility Equation or an isoconversion stability estimate, and release of code and data. None of this requires new methodology.

Fourth, closed-loop platforms that optimise across objectives and scales at once. Present demonstrations address solubility [54], a crystallisation process [55] or tablet specification [56] largely in isolation, whereas the formulation problem is the joint one, and the constrained multi-objective machinery for it already exists [44,46].

11. Conclusion

Descriptor-based prediction brought formulation science a genuine capability and then ran into a wall built from its own reference data. Aqueous solubility models are now accurate to roughly the reproducibility of the measurements that define the problem, and further gains will come from better data rather than better architectures.

Stability prediction has more room, and correspondingly more value, but it lacks the interlaboratory calibration that made the solubility ceiling visible. The most consequential development is not a better predictor at all. It is the reframing of formulation as a sequential decision problem, in which algorithms choose experiments and robots run them, and in which the scarce resource being conserved is drug substance. That reframing has already produced in-specification tablets in hours and high-solubility formulations from a small fraction of a design space.

The technology is ahead of the evidence standards that would let it be trusted. Closing that gap is a matter of reporting discipline, shared data including failures, and orthogonal verification of automated analysis, none of which requires an advance in machine learning.

Declarations

Conflicts of interest: None declared.

Funding: None.

References

1. International Council for Harmonisation. ICH harmonised tripartite guideline Q8(R2): pharmaceutical development. Geneva: ICH; August 2009.
2. International Council for Harmonisation. ICH harmonised tripartite guideline Q1A(R2): stability testing of new drug substances and products. Geneva: ICH; 6 February 2003.
3. Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282-91. doi:10.1016/j.trechm.2019.02.007
4. Abolhasani M, Kumacheva E. The rise of self-driving labs in chemical and materials sciences. Nat Synth. 2023;2(6):483-92. doi:10.1038/s44160-022-00231-0
5. Wang W, Ye Z, Gao H, Ouyang D. Computational pharmaceutics: a new paradigm of drug delivery. J Control Release. 2021;338:119-36. doi:10.1016/j.jconrel.2021.08.030
6. 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
7. Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-driving laboratories for chemistry and materials science. Chem Rev. 2024;124(16):9633-732. doi:10.1021/acs.chemrev.4c00055
8. Jain N, Yalkowsky SH. Estimation of the aqueous solubility I: application to organic nonelectrolytes. J Pharm Sci. 2001;90(2):234-52. doi:10.1002/1520-6017(200102)90:2<234::AID-JPS14>3.0.CO;2-V
9. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci. 2004;44(3):1000-5. doi:10.1021/ci034243x
10. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742-54. doi:10.1021/ci100050t
11. 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):3370-88. doi:10.1021/acs.jcim.9b00237

12. 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):513-30. doi:10.1039/c7sc02664a

13. 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

14. Elbadawi M, Muñiz Castro B, Gavins FKH, Ong JJ, Gaisford S, Pérez G, et al. M3DISEEN: a novel machine learning approach for predicting the 3D printability of medicines. Int J Pharm. 2020;590:119837. doi:10.1016/j.ijpharm.2020.119837

15. 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

16. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform. 2018;10(1):10. doi:10.1186/s13321-018-0263-1

17. Organisation for Economic Co-operation and Development. Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. OECD Series on Testing and Assessment No. 69, ENV/JM/MONO(2007)2. Paris: OECD; 2007.

18. Vermeire FH, Chung Y, Green WH. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. J Am Chem Soc. 2022;144(24):10785-97. doi:10.1021/jacs.2c01768

19. Llinàs A, Glen RC, Goodman JM. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? J Chem Inf Model. 2008;48(7):1289-303. doi:10.1021/ci800058v

20. Llinàs A, Oprisiu I, Avdeef A. Findings of the second challenge to predict aqueous solubility. J Chem Inf Model. 2020;60(10):4791-803. doi:10.1021/acs.jcim.0c00701

21. Palmer DS, Mitchell JBO. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? Mol Pharm. 2014;11(8):2962-72. doi:10.1021/mp500103r

22. Avdeef A. Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. ADMET DMPK. 2019;7(3):210-19. doi:10.5599/admet.698

23. Llompart P, Minoletti C, Baybekov S, Horvath D, Marcou G, Varnek A. Will we ever be able to accurately predict solubility? Sci Data. 2024;11(1):303. doi:10.1038/s41597-024-03105-6

24. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun. 2020;11(1):5753. doi:10.1038/s41467-020-19594-z

25. Francoeur PG, Koes DR. SolTranNet: a machine learning tool for fast aqueous solubility prediction. J Chem Inf Model. 2021;61(6):2530-6. doi:10.1021/acs.jcim.1c00331

26. Sorkun MC, Koelman JMVA, Er S. Pushing the limits of solubility prediction via quality-oriented data selection. iScience. 2021;24(1):101961. doi:10.1016/j.isci.2020.101961

27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. Red Hook (NY): Curran Associates; 2017. p. 4765-74.

28. Baird JA, Van Eerdenbrugh B, Taylor LS. A classification system to assess the crystallization tendency of organic molecules from undercooled melts. J Pharm Sci. 2010;99(9):3787-806. doi:10.1002/jps.22197

29. Alhalaweh A, Alzghoul A, Kaialy W, Mahlin D, Bergström CAS. Computational predictions of glass-forming ability and crystallization tendency of drug molecules. Mol Pharm. 2014;11(9):3123-32. doi:10.1021/mp500303a

30. Dong J, Gao H, Ouyang D. PharmSD: a novel AI-based computational platform for solid dispersion formulation design. Int J Pharm. 2021;604:120705. doi:10.1016/j.ijpharm.2021.120705

31. Lu T, Wu Y, Xiong P, Zhong H, Ding Y, Li H, et al. Combining high-throughput screening and machine learning to predict the formation of both binary and ternary amorphous solid dispersion formulations for early drug discovery and development. Pharm Res. 2025;42(4):697-709. doi:10.1007/s11095-025-03853-z

32. Wicker JGP, Cooper RI. Will it crystallise? Predicting crystallinity of molecular materials. CrystEngComm. 2015;17(9):1927-34. doi:10.1039/C4CE01912A

33. Devogelaer JJ, Meekes H, Tinnemans P, Vlieg E, de Gelder R. Co-crystal prediction by artificial neural networks. Angew Chem Int Ed. 2020;59(48):21711-18. doi:10.1002/anie.202009467

34. Waterman KC, Adami RC. Accelerated aging: prediction of chemical stability of pharmaceuticals. Int J Pharm. 2005;293(1-2):101-25. doi:10.1016/j.ijpharm.2004.12.013

35. Waterman KC. The application of the Accelerated Stability Assessment Program (ASAP) to quality by design (QbD) for drug product stability. AAPS PharmSciTech. 2011;12(3):932-7. doi:10.1208/s12249-011-9657-3

36. Hang NT, Long NT, Duy ND, Chien NN, Phuong NV. Towards safer and efficient formulations: machine learning approaches to predict drug-excipient compatibility. Int J Pharm. 2024;653:123884. doi:10.1016/j.ijpharm.2024.123884

37. Gentiluomo L, Roessner D, Frieß W. Application of machine learning to predict monomer retention of therapeutic proteins after long term storage. Int J Pharm. 2020;577:119039. doi:10.1016/j.ijpharm.2020.119039

38. Narayanan H, Dingfelder F, Condado Morales I, Patel B, Heding KE, Bjelke JR, et al. Design of biopharmaceutical formulations accelerated by machine learning. Mol Pharm. 2021;18(10):3843-53. doi:10.1021/acs.molpharmaceut.1c00469

39. Narayanan H, Dingfelder F, Butté A, Lorenzen N, Sokolov M, Arosio P. Machine learning for biologics: opportunities for protein engineering, developability, and formulation. Trends Pharmacol Sci. 2021;42(3):151-65. doi:10.1016/j.tips.2020.12.004

40. Muñiz Castro B, Elbadawi M, Ong JJ, Pollard T, Song Z, Gaisford S, et al. Machine learning predicts 3D printing performance of over 900 drug delivery systems. J Control Release. 2021;337:530-45. doi:10.1016/j.jconrel.2021.07.046

41. Bannigan P, Bao Z, Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A, et al. Machine learning models to accelerate the design of polymeric long-acting injectables. Nat Commun. 2023;14(1):35. doi:10.1038/s41467-022-35343-w

42. Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the human out of the loop: a review of Bayesian optimization. Proc IEEE. 2016;104(1):148-75. doi:10.1109/JPROC.2015.2494218

43. Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. Phoenics: a Bayesian optimizer for chemistry. ACS Cent Sci. 2018;4(9):1134-45. doi:10.1021/acscentsci.8b00307

44. Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A. Bayesian optimization with known experimental and design constraints for chemistry applications. Digit Discov. 2022;1(5):732-44. doi:10.1039/D2DD00028H

45. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Martinez Alvarado JI, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature. 2021;590(7844):89-96. doi:10.1038/s41586-021-03213-y

46. Cao L, Russo D, Felton K, Salley D, Sharma A, Keenan G, et al. Optimization of formulations using robotic experiments driven by machine learning DoE. Cell Rep Phys Sci. 2021;2(1):100295. doi:10.1016/j.xcrp.2020.100295

47. Tamasi MJ, Patel RA, Borca CH, Kosuri S, Mugnier H, Upadhya R, et al. Machine learning on a robotic platform for the design of polymer-protein hybrids. Adv Mater. 2022;34(30):2201809. doi:10.1002/adma.202201809

48. Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1):5966. doi:10.1038/s41467-020-19597-w

49. Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature. 2020;583(7815):237-41. doi:10.1038/s41586-020-2442-2

50. MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(20):eaaz8867. doi:10.1126/sciadv.aaz8867

51. Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: a community perspective. Matter. 2021;4(9):2702-26. doi:10.1016/j.matt.2021.06.036

52. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0

53. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8

54. Ros H, Abdalla Y, Cook MT, Shorthouse D. Efficient discovery of new medicine formulations using a semi-self-driven robotic formulator. Digit Discov. 2025;4(8):2263-72. doi:10.1039/D5DD00171D

55. Pickles T, Leghrib Y, Weisshaar M, Goncharuk M, Timperman P, Doherty T, et al. Automated scale-up crystallisation DataFactory for model-based pharmaceutical process development: a Bayesian case study. Digit Discov. 2025;4(8):2025-32. doi:10.1039/D4DD00406J

56. Abbas F, Salehian M, Hou P, Moores J, Goldie J, Tsioutsios A, et al. Accelerated drug development using a digital formulator and a self-driving tableting data factory. Nat Commun. 2026;17:4739. doi:10.1038/s41467-026-71204-6

57. Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature. 2023;624(7990):86-91. doi:10.1038/s41586-023-06734-w. Author correction: Nature. 2026;650:E1. doi:10.1038/s41586-025-09992-y

58. Leeman J, Liu Y, Stiles J, Lee SB, Bhatt P, Schoop LM, et al. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. PRX Energy. 2024;3(1):011002. doi:10.1103/PRXEnergy.3.011002

59. 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 and other interested parties. Docket FDA-2024-D-4689. Silver Spring (MD): FDA; 7 January 2025.

60. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Amsterdam: EMA; adopted September 2024.

61. International Council for Harmonisation. ICH harmonised guideline Q1: stability testing of drug substances and drug products, draft version endorsed at Step 2. Geneva: ICH; 11 April 2025.

62. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18