Journal of Advanced Pharmaceutical Sciences and Natural Products

Journal of Advanced Pharmaceutical Sciences and Natural Products (JAPSNP)

MACHINE LEARNING IN PHARMACEUTICAL FORMULATION DEVELOPMENT: FROM MOLECULAR DESCRIPTORS AND PROPERTY PREDICTION TO CLOSED-LOOP AUTONOMOUS EXPERIMENTATION

Mahima Beohar*[1,2]

Dr. Megha Verma[1]

Anupam Jaiswal[1]

Department of Pharmacy, Gyan Ganga Institute of Technology & Sciences, Tilwara Ghat, Jabalpur, Madhya Pradesh, 482003, INDIA

Department of Pharmaceutical Sciences, Shalom Institute of Health and Allied Sciences, Sam Higginbottom University of Agriculture, Technology & Sciences (SHUATS), Naini, Prayagraj, Uttar Pradesh, 211007, INDIA

Corresponding author:
Mahima Beohar

Email: Mahimabeohar@ggits.org

ABSTRACT

Formulation development still consumes a disproportionate share of the time and material spent on a drug candidate, largely because the relationships linking molecular structure, excipient choice, composition and process to product performance are nonlinear and only partly understood. Machine learning (ML) has entered this space along three fronts: numerical representation of molecules and formulations, prediction of developability properties such as solubility and amorphous stability, and, most recently, closed-loop experimentation in which a model chooses the next experiment and a robotic platform performs it. This review traces that progression using only in silico, physicochemical and in vitro evidence, and deliberately excludes studies whose conclusions rest on human or animal data. Across the literature, predictive accuracy for well-posed physicochemical tasks now approaches experimental reproducibility, and dosage-form models trained on a few hundred to several thousand curated records reach useful accuracy for solid dispersions, cyclodextrin complexes, long-acting injectables, printed dosage forms and spectroscopic dissolution testing. The dominant constraint, however, is no longer the algorithm. Formulation data are scarce, heterogeneous, mined from publications that rarely report failure, and routinely evaluated with splits that overstate generalisation. Bayesian optimisation and self-driving laboratories matter chiefly because they change how data are produced: reported campaigns reached target formulations after sampling a few per cent of a design space, or with less than half the experiments of conventional design of experiments. The review argues that the next advance depends on system-level representations of the formulation, standardised and FAIR data including negative results, honest uncertainty estimates, and validation practices aligned with the credibility frameworks now emerging from regulators.

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

1. Introduction

A new chemical entity becomes a medicine only when it is turned into a dosage form that releases its drug reproducibly and stays stable on the shelf. That translation has long been an empirical craft. Pre-formulation studies characterise the molecule; a formulator then selects excipients and process conditions from experience, runs a screening design, and iterates. Quality by design, as codified in ICH Q8(R2), gave the craft a statistical grammar by asking developers to define a design space, the multidimensional combination of material attributes and process parameters shown to assure quality [1]. What the guideline could not supply was a means of predicting where that space lies before the experiments are run.

Machine learning offers such a means, at least in principle. The appeal is straightforward. Formulation problems are high-dimensional, response surfaces are rugged, and classical response-surface designs scale poorly once more than a handful of factors interact. A model that learns the mapping from inputs to quality attributes could shortlist candidates, flag unstable compositions and reduce the experimental burden [2,3]. Over the past decade the field has moved through three recognisable stages. The first concerned representation: how to turn a molecule, and later an entire formulation, into numbers a learning algorithm can use. The second concerned prediction of individual properties and then of dosage-form performance. The third, still young, couples prediction to automated experimentation so that the model and the laboratory improve each other in a closed loop [4].

This review follows that arc, but with a particular argument in view. Published accounts tend to present the field as algorithm-limited, as if the next architecture will unlock formulation design. The evidence assembled here points elsewhere. Where data are abundant and endpoints well defined, current methods already approach the noise floor of the measurement. Where they fail, the cause is almost always the data: too few records, inconsistent reporting, missing failures, and evaluation schemes that reward memorisation. Read this way, the real contribution of closed-loop experimentation is less the robot than the dataset it leaves behind.

2. Scope and approach

This is a narrative review. Literature was identified through PubMed and Google Scholar, supplemented by publisher databases, using combinations of the terms machine learning, deep learning, formulation, solubility, solid dispersion, dissolution, Bayesian optimisation, active learning and self-driving laboratory, with snowballing from recent reviews. Regulatory documents were retrieved directly from the agencies that issued them. Priority was given to studies with prospective experimental confirmation and to benchmark papers that define the state of the art. In keeping with the brief, studies whose principal findings depend on human participants or live animals, including in vivo pharmacokinetic, efficacy or transfection outcomes, were excluded; where a paper

combined in vitro and in vivo work it was omitted rather than cited selectively. The included studies report different endpoints (classification accuracy, coefficients of determination, mean absolute errors, dissolution similarity factors, experimental savings) on unrelated datasets. They do not estimate a common quantity, so no quantitative pooling was attempted, and the synthesis is structured narratively around Table 2.

3. Representing a formulation

3.1 Molecular descriptors and fingerprints

Every model begins with a representation, and in formulation science the molecule was the natural starting point. Classical descriptors encode physicochemical and topological features: molecular weight, calculated logP, polar surface area, hydrogen-bond counts, connectivity indices. Open tools have made such calculation routine; Mordred, for example, computes more than 1800 two- and three-dimensional descriptors and runs at least twice as fast as the older PaDEL package [5]. Extended-connectivity fingerprints take a different route, hashing circular atom environments into a bit vector that is not restricted to a predefined dictionary of substructures [6]. Descriptors carry physical meaning and are easy to interpret; fingerprints capture local structure but say little about bulk behaviour. For formulation endpoints governed by crystal packing or melt thermodynamics, that distinction matters, as Section 4 shows.

3.2 Learned representations

Graph neural networks learn their own features from the molecular graph. The directed message-passing network introduced by Yang and colleagues, later released as Chemprop, matched or outperformed fixed descriptors on 19 public and 16 proprietary industrial datasets, although the authors were careful to note that accuracy had not yet reached experimental reproducibility [7]. The MoleculeNet benchmark reached a similar verdict from the other direction: learnable representations generally won, but struggled on small and imbalanced datasets, and for some tasks physics-aware featurisation mattered more than the choice of algorithm [8]. Chemical language models now extend the idea by pretraining on unlabelled structures at scale; MoLFormer was trained on 1.1 billion molecules from PubChem and ZINC [9]. General-purpose large language models, fine-tuned on a few dozen examples, have matched specialist models in some low-data chemical tasks [10]. That last finding deserves attention from formulators, whose datasets rarely exceed a few hundred records.

3.3 The missing layer

A formulation is not a molecule. It is a drug, one or more excipients in defined proportions, and a manufacturing history. The encodings described above address only the first element. Published dosage-form models have represented the remainder in improvised ways: excipient identities as one-hot categories or trade names, compositions as weight fractions, process settings as scalar inputs, and the polymer often as

a handful of bulk properties [11,12]. None of these is transferable. A model that has learned what hydroxypropyl methylcellulose does in one matrix cannot reason about a chemically similar grade it has never seen, because the grade was never described chemically. System-level representation, in which excipients, their ratios and the process are encoded on a common physical footing, is arguably the least developed and most consequential problem in the whole field. Table 1 sets out what each family of representation captures and where it breaks down.

Table 1. Families of representation used in formulation machine learning, the information each encodes, and its principal limitation for dosage-form problems. The assessment synthesises references [5-9,11-13].

Representation	What it encodes	Strength	Principal limitation in formulation
Computed descriptors (e.g., Mordred)	Global physicochemical and topological properties of the drug	Physically interpretable; works with small data	Ignores excipients, solid form and process
Circular fingerprints (ECFP)	Presence of local atom environments	Fast; strong baseline for structure-based tasks	Weak link to bulk properties such as lattice energy
Learned graph embeddings	Task-specific features from the molecular graph	Often best when data are plentiful	Data-hungry; unstable on a few hundred records
Pretrained chemical language models	Statistical patterns from very large unlabelled corpora	Transfer to low-data tasks	Opaque; pretraining corpus contains no formulation information
Composition and process vectors	Excipient identities, ratios and process settings	Captures the formulation as made	Excipients encoded by name, so no transfer to unseen grades
Physics-informed hybrids	Learned terms inside a thermodynamic or kinetic framework	Extrapolates beyond training conditions	Requires a valid mechanistic scaffold for each endpoint

4. Property prediction as the entry point

4.1 Solubility: an instructive ceiling

Aqueous solubility is the property most often modelled and the one that best exposes the limits of prediction. Delaney's ESOL, a linear model built on 2874 measurements and nine molecular properties, predicted within a factor of roughly five to eight of measured values and set an early baseline [14]. Curation later became as important as modelling. AqSolDB merged nine public sources into 9982 unique compounds with reliability labels, a recognition that inconsistent entries in legacy datasets had been silently degrading every model trained on them [15].

The two Solubility Challenges made the ceiling explicit. When organisers assembled test sets from multilaboratory shake-flask data, the interlaboratory standard deviation was about 0.17 log units for a tight set and about 0.62 log units for a set of contentious compounds [16,17]. No model can be validated below the noise of its reference values. Against this background, Boobier and co-workers combined several learners with computed physicochemical terms and placed 79.8% of predictions within ±0.7 log units for their best solvent set, accuracy close to typical experimental uncertainty [18]. Vermeire and colleagues embedded learned solvation free energies and enthalpies within a thermodynamic cycle, extending prediction across solvents and

to temperatures of 550 K or higher, with validation against more than 5000 experimental values [13]. The lesson is general. Solubility of a crystalline solid depends on lattice energy as well as solvation, so purely structural models reach a plateau that hybrid, physics-informed models can partly break through.

4.2 Solid-state behaviour

Amorphous formulations raise a harder question: will the drug stay amorphous? Baird, Van Eerdenbrugh and Taylor screened 51 organic molecules by heat-cool-heat calorimetry and sorted them into three crystallisation-tendency classes, linking poor glass-forming ability to low molecular weight and rigid structure [19]. Alhalaweh and co-workers extended the work to 131 drugs and support vector machines, confirming a molecular-weight threshold near 300 and identifying a single three-dimensional descriptor that classified 86% of compounds correctly irrespective of molecular weight [20]. These studies illustrate a productive pattern: a physically meaningful, reproducible in vitro assay generates a consistent dataset, and modest models then perform well.

Crystal structure prediction sits at the far end of the difficulty scale. The seventh blind test, organised by the Cambridge Crystallographic Data Centre, included a flexible polymorphic drug candidate, a cocrystal and a polymorphic salt; many groups solved the small rigid target, while few succeeded on the complex systems [21]. Polymorph risk therefore remains partly an experimental matter, and any formulation model that ignores solid form inherits that uncertainty.

5. From property to product: machine learning for dosage-form design

Table 2 summarises representative dosage-form studies whose conclusions rest on in silico or in vitro data. Three features stand out. Dataset sizes are small by ML standards, from 73 manufacturing batches to roughly 8000 release measurements. Tree ensembles, particularly random forests and gradient-boosted models, dominate. And prospective confirmation, although increasingly common, remains the exception rather than the rule [2].

Table 2. Representative machine-learning studies in formulation design and pharmaceutical processing that rely on in silico or in vitro data only. Dataset sizes and performance metrics are reported as given by the original authors and are not comparable across rows because endpoints and validation schemes differ from study to study.

System and task	Data	Best model	Reported performance	Ref.
Oral films and sustained-release matrix tablets: disintegration and dissolution class	131 film and 145 tablet formulations from literature	DNN	about 80% accuracy on validation and test sets	[22]
Amorphous solid dispersions: physical stability	646 stability records, 50 drugs, 25 polymers	RF	82.5% test accuracy; prospective check with a 17β-estradiol and PVP dispersion	[23,24]
Cyclodextrin inclusion complexes: binding free energy	about 3000 records, 1320 guests, 8 cyclodextrins	LightGBM	test R^2 0.86, MAE 1.38 kJ/mol; R^2 fell to 0.58 when trained on 500 records	[25]
PLGA long-acting injectables: in vitro release	curated release data, 8013 instances in test folds	LightGBM	model-guided microparticles for two drugs confirmed in vitro	[12]
FDM-printed dosage forms: printability and process	614 formulations, 145 excipients	several ML algorithms	76% accuracy for printability; MAE 8.3 °C for printing	[11]

temperature			temperature	
Printed dosage forms: extrusion, printing and release	968 formulations from 114 articles	several; ANN for release	up to 93% accuracy; mean error of about 24 min for release time points	[26]
Extended-release HPMC tablets: dissolution profile from spectra	148 tablets, 37 manufacturing settings, NIR and Raman	ANN	all test predictions within f2 of 50 to 100	[27]
Commercial wet granulation: granule size and tablet tensile strength	73 production batches, 2017 to 2020	ridge regression, RF	product-specific critical factors identified retrospectively	[28]

5.1 Solid dispersions and complexation

Amorphous solid dispersions are a natural target because stability failure is costly and slow to observe. Han and colleagues compiled 646 stability records, compared eight algorithms, and found random forests the most accurate at 82.5%, then confirmed a prediction experimentally with a new dispersion [23]. The same group converted these models into a public web platform, PharmSD, that predicts stability and dissolution behaviour from the drug structure alone [24]. The convenience is real. So is the limitation, since a polymer described only by identity cannot inform predictions for a polymer outside the training set.

Cyclodextrin complexation offers a cleaner test of data sufficiency. With about 3000 records, gradient boosting reached a test R^2 of 0.86; truncating the training set to 500 records dropped it to 0.58 [25]. Few formulation papers report such a learning curve, yet it answers the practical question every development team asks: how much data is enough?

5.2 Release-controlled and long-acting systems

Release from polymeric carriers depends on drug, polymer, loading, particle size and the release medium, a combination that resists mechanistic modelling. Yang and co-workers showed early that deep networks could classify disintegration and dissolution outcomes for oral films and sustained-release tablets with roughly 80% accuracy, and introduced a dissimilarity-based data-splitting algorithm to cope with small, imbalanced sets [22]. Bannigan and colleagues then curated in vitro release data for long-acting injectables, found gradient-boosted trees the strongest of the models tested, and used the final model to design PLGA microparticles whose release was confirmed experimentally [12]. That study remains one of the clearest demonstrations that a formulation model can guide new work rather than merely fit old data.

5.3 Printed dosage forms

Three-dimensional printing of medicines produces unusually structured data: each formulation has a filament, a printing temperature and a pass or fail outcome. The M3DISEEN platform used 614 fused-deposition formulations to predict printability with 76% accuracy and printing temperature with a mean absolute error of 8.3 °C [11]. A follow-up drew 968 formulations from 114 publications and reached accuracies up to 93% for some extrusion and printing outcomes [26]. The authors themselves identified scarce training data as the main barrier [29], and literature-mined datasets carry a bias worth naming: failed prints are seldom published.

5.4 Process understanding and real-time quality

At the manufacturing end, ML serves as a soft sensor. Galata and co-workers predicted full dissolution profiles of extended-release tablets from near-infrared and Raman spectra, with every test prediction inside the f2 acceptance range and neural networks outperforming partial least squares [27]. A related study on caffeine tablets reported the best performance from reflection near-infrared spectra, with a mean f2 of 63 [30]. Machine vision now complements spectroscopy in granulation and coating, and increasingly in dissolution monitoring [31]. Mäki-Lohiluoma and colleagues showed that 73 routine production batches could yield product-specific process understanding retrospectively, effectively retro-fitting quality-by-design insight onto a legacy product [28]. These applications sit closest to regulatory use, a point taken up in Section 8.

6. Closing the loop

6.1 Bayesian optimisation as the decision engine

Predictive models answer the question of what a given formulation will do. Development asks a different question: which formulation should be made next? Bayesian optimisation addresses the second directly. A probabilistic surrogate, typically a Gaussian process, approximates the response surface and reports its own uncertainty; an acquisition function then balances exploitation of promising regions against exploration of uncertain ones [32]. Chemistry-specific optimisers such as Phoenics adapted the approach to experimental noise and parallel batches [33]. In a direct comparison built around reaction optimisation, a Bayesian optimiser outperformed expert chemists on average efficiency and consistency [34]. Figure 1 places this decision engine within the full closed loop.

Formulation adds two complications that reaction optimisation largely escapes. Objectives are rarely single: a formulator wants solubility, physical stability, viscosity and manufacturability at once, so the useful output is a Pareto front of trade-offs rather than one optimum. Constraints are also hard. An excipient level above its accepted limit, or a composition that cannot be processed, is not a poor result but an inadmissible one, and the optimiser must know that before it proposes the experiment. Multi-objective and constrained variants of Bayesian optimisation address both issues, at the price of more careful specification by the scientist at the outset.

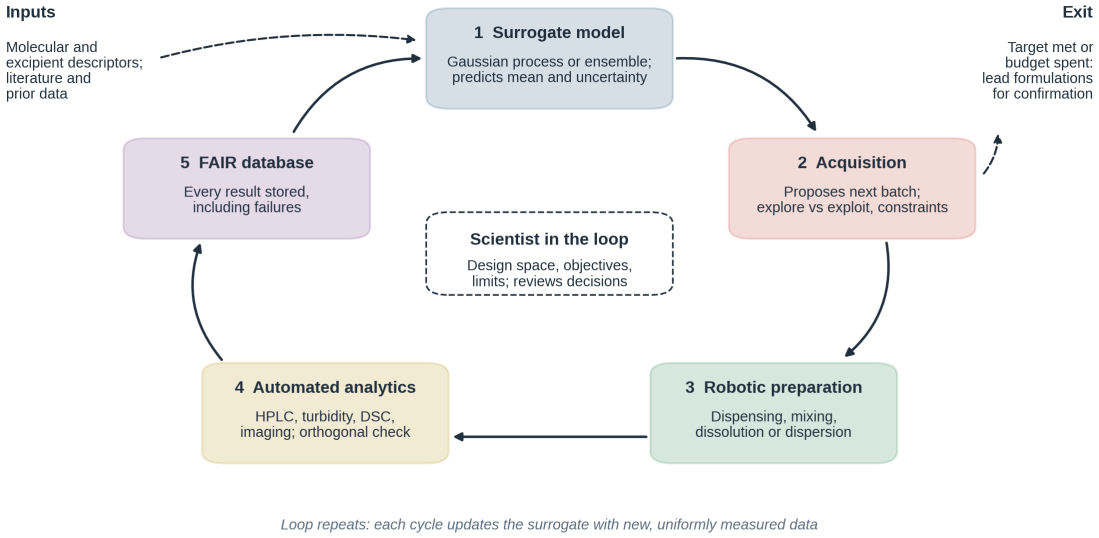

Figure 1. Architecture of a closed-loop formulation laboratory. Molecular and excipient descriptors, together with prior data, initialise a probabilistic surrogate model. An acquisition function proposes the next batch of formulations, which a robotic platform prepares and characterises with automated analytics such as HPLC, turbidity or imaging. Results, including failures, are written to a FAIR database and used to update the surrogate, and the cycle repeats until the quality target or the experimental budget is reached. The scientist defines the design space, constraints and objectives, and reviews each decision point. Schematic synthesised from the architectures described in references [4,35-37]. FAIR, findable, accessible, interoperable and reusable; HPLC, high-performance liquid chromatography.

6.2 Lessons from chemistry and materials

The self-driving laboratory matured first outside pharmaceutics. A mobile robotic chemist worked autonomously for eight days, completed 688 experiments in a ten-variable space and found photocatalyst mixtures about six times more active than the starting formulations [35]. The Ada platform optimised organic thin films for hole mobility in closed loop [38]. These studies established that automation and Bayesian search can be integrated robustly for days at a time.

They also supplied a cautionary tale. The A-Lab reported the synthesis of 41 new inorganic compounds from 58 targets in 17 days [39]. Independent re-analysis questioned the automated interpretation of the diffraction data and argued that many claimed successes were likely known disordered phases [40]. A published author correction subsequently reduced the confirmed count and clarified that the materials were new to the prediction platform rather than necessarily new to science [41]. The episode is directly relevant to formulation. A closed loop is only as reliable as the automated measurement that closes it, and an unvalidated analytical step can generate confident, wrong conclusions at robotic speed.

6.3 Evidence from formulation

Closed-loop work on medicines is recent but consistent in direction (Figure 2). Cao and colleagues combined classification with Thompson sampling on semi-automated robots and found nine liquid formulations meeting target criteria within 15 working days, albeit for

a consumer product rather than a drug [42]. For biopharmaceuticals, Narayanan and co-workers identified the formulation giving maximal thermal stability for three antibody-fragment variants within 25 experiments, less than one third of the effort of a classical design [43]. Adaptive Bayesian optimisation of cooling crystallisation cut the number of experiments by 54% for lamivudine and 79% for aspirin, relative to design of experiments [44], and the same group has extended the approach to automated scale-up with in-line analytics [45].

Two studies from a single group sharpen the picture. A semi-self-driven robotic formulator searching for curcumin solubilising systems sampled 256 of 7776 candidate formulations, about 3.3% of the space, and found seven leads above 10 mg/mL in six days [37]. Multi-objective optimisation of a thermoresponsive nasal gel then reached acceptable trade-offs among several attributes after preparing only 22 samples; nine of eleven model-guided formulations met the gelation window of 28 to 36 °C, against three of eight initial designs [46]. Hickman and colleagues have proposed a roadmap for extending such platforms to nanomedicines, again identifying consistent data as the bottleneck [36].

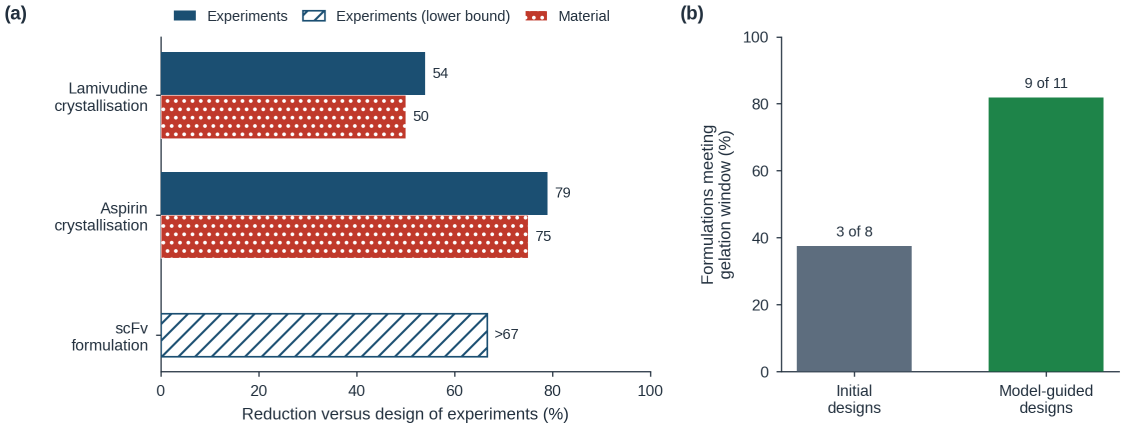

Figure 2. Closed-loop and Bayesian-optimisation campaigns reduced experimental effort and raised the hit rate relative to conventional designs. (a) Reduction in the number of experiments (solid bars) and in material consumed (dotted bars) relative to design of experiments, as stated by the original authors for batch cooling crystallisation of lamivudine and aspirin; the hatched bar for single-chain variable fragment (scFv) formulation is a lower bound, because the source reports that the optimum was found with less than one third of the experiments of a classical design, and no material figure was reported. (b) Proportion of thermoresponsive nasal-gel formulations meeting the target gelation window of 28 to 36 °C among the initial designs (3 of 8) and the formulations proposed by multi-objective Bayesian optimisation (9 of 11). Each bar is a single reported value without dispersion, so no error bars are shown and no statistical test was applied. Values are taken directly from references [43,44,46] and are not pooled, because the studies differ in objective and in baseline.

These figures should be read with care. Each study defined its own baseline, and a percentage saving against one design of experiments says little about savings against a better one. What they establish is that sequential, model-guided sampling is consistently more economical than one-shot designs for problems of moderate dimension. They also produce something one-

shot designs rarely do: a dataset in which every attempt, successful or not, is recorded under uniform conditions.

6.4 Language-model agents

Large language models have begun to act as planners. Coscientist, a GPT-4-based agent with web search, code execution and robot control, planned and executed tasks including optimisation of palladium-catalysed couplings [47]. ChemCrow gave a language model 18 expert tools and used it to plan syntheses and guide discovery of a new chromophore [48]. For formulation, the attraction lies in retrieving literature precedent and translating a development goal into an experimental plan. The risk is equally clear. A fluent plan is not a validated one, and such agents belong inside a loop with human review and verified analytics, not in place of them.

7. Critical appraisal: why models fail to travel

The formulation ML literature contains many high reported accuracies and comparatively few models in routine use. Several recurring weaknesses explain the gap.

The first is the data themselves. Most dosage-form datasets are assembled from published papers, which differ in method and reporting and almost never include failed formulations. Raccuglia and co-workers showed, in inorganic synthesis, that archived failed experiments allowed a model to predict successful conditions with 89% success, outperforming human strategies [49]. The formulation field has no equivalent archive. A community checklist for chemical ML also warns that public databases commonly contain more than 10% erroneous entries [50].

The second is evaluation. Random splits place near-duplicate formulations, often the same drug and polymer at different ratios, in both training and test sets. Wallach and Heifets demonstrated that benchmark performance in ligand-based modelling tracks such train-test overlap closely, whatever the method [51]. Formulation studies rarely report splits by drug, by polymer or by time, which are the splits that mimic real use. Until they do, reported accuracies should be regarded as upper bounds.

Uncertainty is the third. A model that cannot say when it is extrapolating is dangerous in development. Applicability-domain methods based on similarity to the training set and on ensemble spread were described for random forests more than a decade ago [52]. For neural networks, a systematic comparison found no uncertainty method uniformly superior across datasets [53]. Closed-loop optimisation depends on these estimates, so their calibration deserves the same scrutiny as point accuracy.

Interpretability is the fourth. Attribution methods such as SHAP apportion a prediction among its inputs on a consistent theoretical basis [54], and they are now widely applied to formulation models. They explain what a model has learned, not what is physically true, and correlated inputs such as drug loading and polymer ratio can produce attributions that mislead. Mechanistic

hypotheses drawn from attribution plots need experimental testing.

Finally, benchmarks are missing. Molecular property prediction has MoleculeNet and the Solubility Challenges [8,17]; formulation has no shared, curated test sets with defined splits. The consequence is a literature of individually persuasive studies that cannot be compared.

Weighed together, the evidence supports a graded verdict. It is strong for single physicochemical properties measured by standardised assays, where curated datasets of thousands of compounds exist and error is bounded by measurement noise. It is moderate for dosage-form classification and regression tasks, where models trained on a few hundred records perform well within their own data but have seldom been tested on drugs or excipients from outside it. It is still thin for autonomous formulation, where the published campaigns are few, mostly from a small number of groups, and optimise objectives chosen by the same investigators who report the savings. None of this diminishes the progress. It does locate the next task precisely: independent replication, external test sets and prospective comparison against well-designed conventional experiments.

8. Regulatory and quality considerations

Regulators have moved from observation to draft policy. The FDA's 2023 discussion paper on artificial intelligence in drug manufacturing sought comment on uses such as process design, advanced control and monitoring for products under new and abbreviated applications [55]. Its January 2025 draft guidance proposes a risk-based credibility assessment built around the model's context of use, and covers nonclinical and manufacturing applications while excluding drug discovery [56]. The EMA reflection paper, published in September 2024, addresses artificial intelligence across the medicinal product lifecycle [57]. ICH Q14 already anticipates model lifecycle management, describing monitoring of multivariate calibration models that may trigger recalibration [58].

For formulation scientists, the practical implication is that model credibility will be judged against intended use. A model that ranks candidates for early screening carries little regulatory risk; a model that supports a design space or real-time release carries a great deal. Documentation of where the training data came from, how the model was validated, and how drift will be detected will be expected in proportion. Closed-loop platforms fit well with this philosophy, provided their data management follows the FAIR principles, findable, accessible, interoperable and reusable, which were written with machine use explicitly in mind [59].

9. Future directions

Several specific developments would move the field forward faster than any new architecture.

Shared formulation benchmarks. Curated datasets for amorphous stability, release and printability, with fixed splits by drug and by excipient, would allow methods to be compared honestly and would reveal how far current accuracies depend on leakage.

Excipient representation. Polymers and surfactants need descriptors grounded in chemistry and physical properties such as molecular weight distribution, substitution pattern and glass transition, so that models can generalise across grades and suppliers.

Negative-data reporting. Journals and consortia could require deposition of failed formulations alongside successful ones. Closed-loop platforms generate such data by default and could seed these repositories.

Physics-informed hybrids. The solubility literature shows that embedding thermodynamic structure extends accuracy beyond what purely statistical models achieve [13]. Similar hybrids for dissolution, crystallisation and phase separation are within reach.

Validated autonomous analytics. The A-Lab experience argues for routine orthogonal confirmation of automated measurements before closed-loop conclusions are reported [40,41].

Low-data learning. Fine-tuned language models and transfer from large molecular corpora may help where records number in the dozens [9,10], but they need testing against the simple baselines that often win on small tabular data.

10. Conclusions

Machine learning in formulation has progressed from describing molecules to predicting properties, then dosage-form behaviour, and now to deciding which experiment to run next. For well-defined physicochemical endpoints, prediction already approaches the reproducibility of the measurement. For dosage-form design, useful models exist but are constrained by small, literature-mined, success-biased datasets and by validation that flatters them. Closed-loop experimentation is the most promising response, less because it automates the laboratory than because it produces the systematic, complete and well-annotated data on which every other advance depends.

References

1. International Council for Harmonisation. ICH harmonised tripartite guideline Q8(R2): Pharmaceutical development. Step 4 version. Geneva: ICH; August 2009. Available from: https://database.ich.org/sites/default/files/Q8_R2_Guideline.pdf
2. 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
3. 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
4. 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
5. 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

6. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742-54. doi:10.1021/ci100050t
7. 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
8. 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
9. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
10. Jablonka KM, Schwaller P, Ortega-Guerrero A, Smit B. Leveraging large language models for predictive chemistry. Nat Mach Intell. 2024;6:161-9. doi:10.1038/s42256-023-00788-1
11. 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
12. 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;14:35. doi:10.1038/s41467-022-35343-w
13. 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
14. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci. 2004;44(3):1000-5. doi:10.1021/ci034243x
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. Llinàs A, Avdeef A. Solubility Challenge revisited after ten years, with multilab shake-flask data, using tight (SD ~0.17 log) and loose (SD ~0.62 log) test sets. J Chem Inf Model. 2019;59(6):3036-40. doi:10.1021/acs.jcim.9b00345
17. 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
18. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical

relationships: solubility prediction in organic solvents and water. Nat Commun. 2020;11:5753. doi:10.1038/s41467-020-19594-z

19. 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
20. 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
21. Hunnisett LM, Nyman J, Francia N, Abraham NS, Adjiman CS, Aitipamula S, et al. The seventh blind test of crystal structure prediction: structure generation methods. Acta Crystallogr B Struct Sci Cryst Eng Mater. 2024;80(6):517-47. doi:10.1107/S2052520624007492
22. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharm Sin B. 2019;9(1):177-85. doi:10.1016/j.apsb.2018.09.010
23. 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
24. 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
25. Zhao Q, Ye Z, Su Y, Ouyang D. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. Acta Pharm Sin B. 2019;9(6):1241-52. doi:10.1016/j.apsb.2019.04.004
26. 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
27. Galata DL, Farkas A, Könyves Z, Mészáros LA, Szabó E, Csontos I, et al. Fast, spectroscopy-based prediction of in vitro dissolution profile of extended release tablets using artificial neural networks. Pharmaceutics. 2019;11(8):400. doi:10.3390/pharmaceutics11080400
28. Mäki-Lohiluoma E, Säkkinen N, Palomäki M, Winberg O, Ta HX, Heikkinen T, et al. Use of machine learning in prediction of granule particle size distribution and tablet tensile strength in commercial pharmaceutical manufacturing. Int J Pharm. 2021;609:121146. doi:10.1016/j.ijpharm.2021.121146

29. Elbadawi M, McCoubrey LE, Gavins FKH, Ong JJ, Goyanes A, Gaisford S, et al. Disrupting 3D printing of medicines with machine learning. Trends Pharmacol Sci. 2021;42(9):745-57. doi:10.1016/j.tips.2021.06.002

30. Nagy B, Petra D, Galata DL, Démuth B, Borbás E, Marosi G, et al. Application of artificial neural networks for Process Analytical Technology-based dissolution testing. Int J Pharm. 2019;567:118464. doi:10.1016/j.ijpharm.2019.118464

31. Galata DL, Mészáros LA, Kállai-Szabó N, Szabó E, Pataki H, Marosi G, et al. Applications of machine vision in pharmaceutical technology: a review. Eur J Pharm Sci. 2021;159:105717. doi:10.1016/j.ejps.2021.105717

32. 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

33. 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

34. 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

35. 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

36. Hickman RJ, Bannigan P, Bao Z, Aspuru-Guzik A, Allen C. Self-driving laboratories: a paradigm shift in nanomedicine development. Matter. 2023;6(4):1071-81. doi:10.1016/j.matt.2023.02.007

37. 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

38. 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

39. Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature. 2023;624(7990):86-91. doi:10.1038/s41586-023-06734-w

40. 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

41. Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. Author Correction:

An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature. 2026;650:E1. doi:10.1038/s41586-025-09992-y

42. 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
43. 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
44. Pickles T, Mustoe C, Brown CJ, Florence AJ. Comparative study on adaptive Bayesian optimization for batch cooling crystallization for slow and fast kinetic regimes. Cryst Growth Des. 2024;24(3):1245-53. doi:10.1021/acs.cgd.3c01225
45. 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
46. Liu H, Gucic A, Liu H, Cook MT, Shorthouse D. Fast-tracking complex formulation development with multi-objective Bayesian optimisation. J Control Release. 2026;397:115171. doi:10.1016/j.jconrel.2026.115171
47. 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
48. 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
49. Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature. 2016;533(7601):73-6. doi:10.1038/nature17439
50. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
51. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
52. Sheridan RP. Three useful dimensions for domain applicability in QSAR models using random forest. J Chem Inf Model. 2012;52(3):814-23. doi:10.1021/ci300004n

53. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502

54. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Red Hook (NY): Curran Associates; 2017. Available from: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

55. US Food and Drug Administration. Artificial intelligence in drug manufacturing: discussion paper; notice and request for comments. Fed Regist. 2023 Mar 1;88(40):12943. Docket FDA-2023-N-0487. Available from: https://www.fda.gov/media/165743/download

56. 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. Silver Spring (MD): FDA; January 2025. Docket FDA-2024-D-4689. Available from: https://www.federalregister.gov/documents/2025/01/07/2024-31542

57. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Amsterdam: EMA; 30 September 2024. Available from: https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle-scientific-guideline

58. International Council for Harmonisation. ICH harmonised guideline Q14: Analytical procedure development. Step 4. Geneva: ICH; 1 November 2023. Available from: https://database.ich.org/sites/default/files/ICH_Q14_Guideline_2023_1116_1.pdf

59. 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