



Data-Driven Prediction of Pharmaceutical Solid Forms Across Polymorphism, Crystallization, Molecular Packing, Stability, and Developability Decisions

Laura Keller^{1*}, Thomas Lehmann¹, Simon Brunner², Christoph Meier³, David Thompson⁴

¹Department of Solid-Form Prediction and Polymorphism, Faculty of Pharmacy, ETH Zurich, Zurich, Switzerland.

²Department of Crystallization and Molecular Packing, Faculty of Pharmacy, EPFL Lausanne, Lausanne, Switzerland.

³Department of Stability and Developability, Faculty of Pharmacy, University of Bern, Bern, Switzerland.

⁴Department of Data-Driven Solid-Form Decisions, Faculty of Pharmacy, University of Basel, Basel, Switzerland.

ABSTRACT

Pharmaceutical solid-form development increasingly relies on data-driven and physics-informed models to anticipate crystal structures, polymorph landscapes, multicomponent forms, crystallization behavior, stability, and downstream developability, yet these tasks are often discussed together despite representing different prediction targets, evidence levels, and decision consequences. This systematic review evaluates how current methods connect molecular or process data to solid-state predictions and how far those predictions can support pharmaceutical decisions. A structured search and screening strategy was applied, covering molecular-crystal machine learning, crystal structure prediction, polymorphism, solvates, salts, cocrystals, nucleation, stability, phase transformation, crystal habit, and developability. Study information was coded by material class, target, representation, algorithm, validation design, experimental confirmation, uncertainty source, and transferability, while risk-of-bias appraisal adapted general prediction-model principles with crystal-specific concerns including structural leakage, generated versus experimental labels, split strategy, condition dependence, and external validation. Heterogeneous endpoints were synthesized without indiscriminate quantitative pooling. The literature resolves into distinct layers: solid-form propensity classification, crystal packing and polymorph ranking, process and nucleation modeling, finite-temperature or environment-dependent stability, dynamic transformation modeling, and developability-oriented decision support. The strongest evidence connects computational prediction to blind, external, or experimental validation, whereas major weaknesses arise when model confidence, benchmark accuracy, computed stability, or packing plausibility is treated as equivalent to experimentally realized form behavior. Data-driven solid-form prediction is therefore most defensible as a staged evidence problem rather than a single performance problem, and a proposed task-to-decision hierarchy is used to preserve boundaries between prediction, physical validation, experimental realization, and pharmaceutical consequence.

Key Words: *Pharmaceutical solid forms, Polymorphism, Crystal structure prediction, Machine learning, Crystallization, Developability*

eIJPPR 2026; 16(4):20-29

HOW TO CITE THIS ARTICLE: Keller L, Lehmann T, Brunner S, Meier C, Thompson D. Data-Driven Prediction of Pharmaceutical Solid Forms Across Polymorphism, Crystallization, Molecular Packing, Stability, and Developability Decisions. Int J Pharm Phytopharmacol Res. 2026;16(4):20-9. <https://doi.org/10.51847/Cn0PHb2hHw>

INTRODUCTION

The pharmaceutical relevance of a solid-state model is determined not only by whether it predicts a structural

label, but by whether that prediction corresponds to a decision that matters in development. Recent digital-design work connects molecular structure, polymorph landscapes, and crystallization-process reasoning, while

Corresponding author: Laura Keller

Address: Department of Solid-Form Prediction and Polymorphism, Faculty of Pharmacy, ETH Zurich, Zurich, Switzerland.

E-mail: ✉laura.keller@pharma.ethz.ch

Received: 01 April 2026; **Revised:** 08 July 2026; **Accepted:** 11 July 2026

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



large-scale crystal structure prediction has increasingly been evaluated through broader validation and blind testing rather than isolated retrospective examples [1, 2]. This evolution creates an opportunity, but also a risk of collapsing conceptually different tasks into a common language of “prediction.”

For an active pharmaceutical ingredient, a model may estimate packing, rank candidate polymorphs, classify solvate or cocrystal propensity, approximate nucleation behavior, simulate finite-temperature stability, or guide morphology and coformer choices. These outputs differ in physical meaning, validation requirements, and proximity to a developability decision. A high-ranking hypothetical packing does not establish that the form will nucleate; a predicted stable form does not establish kinetic persistence; and a favorable multicomponent form does not establish a robust drug product. This review therefore treats data-driven solid-form prediction as a layered evidence problem in which computational accuracy, physical plausibility, experimental realization, and decision relevance must be evaluated separately. The central interpretive problem is not simply whether a model is accurate on its stated benchmark, but whether the benchmark, target definition, environmental assumptions, and validation design are appropriate to the inference subsequently made. Accordingly, the review follows evidence across layers without assuming that information gained at an upstream computational layer automatically transfers to a downstream pharmaceutical decision.

Review objective and questions

Recent reviews show that machine learning is being applied both to crystal behavior and crystallization-process control, while pharmaceutical crystallization studies extend AI toward prediction, optimization, and operational decision support [3, 4]. The objective of this review is to determine how these heterogeneous approaches should be interpreted together without treating unlike prediction targets as interchangeable.

The review asks six linked questions. First, what solid-state or process quantity is being predicted: form propensity, space group, packing, polymorph rank, nucleation behavior, stability, transformation, habit, or a downstream property? Second, what data and representation support the prediction? Third, at what level does the model operate, from molecule or molecular pair to crystal structure,

atomistic trajectory, or process state? Fourth, how is performance validated, including the use of internal, grouped, external, blind, or experimental tests? Fifth, which environmental or process conditions bound the inference? Sixth, what pharmaceutical decision can the result legitimately inform?

The organizing contribution proposed here is a task-to-decision interpretation in which evidence is evaluated according to the distance between a computational output and the pharmaceutical consequence inferred from it. The review does not assume that progress at one layer automatically validates another.

Protocol and prisma 2020 reporting framework

The review was structured as a systematic review and reported in accordance with PRISMA 2020 principles, which require transparent description of eligibility, information sources, selection, and synthesis rather than treating the reporting framework itself as evidence of study quality [5]. No claim of prospective protocol registration is made because no verified registration record was available for this review. Before narrative synthesis, the review logic fixed the eligibility boundaries, extraction domains, first-level study classifications, appraisal dimensions, and the distinction between evidence-supported findings and proposed interpretive synthesis. This procedural separation was used to reduce post hoc redefinition of prediction tasks after individual study results were known.

Search reporting was additionally guided by PRISMA-S principles emphasizing reproducible description of sources, search concepts, dates, and retrieval procedures [6]. The preserved search record supported verification of the curated candidate set, individual references, exclusion rationales, and the final evidence set, but did not preserve reproducible source-level raw retrieval, deduplication, title/abstract screening, or full-text-request counts. Those upstream quantities were therefore not reconstructed retrospectively.

This distinction is methodologically important. A visually complete flow diagram can create false precision if counts are inferred after the fact. The reporting strategy therefore separates verified selection information from unavailable database-level accounting. **Figure 1** presents that evidence-selection record without filling unobserved stages.

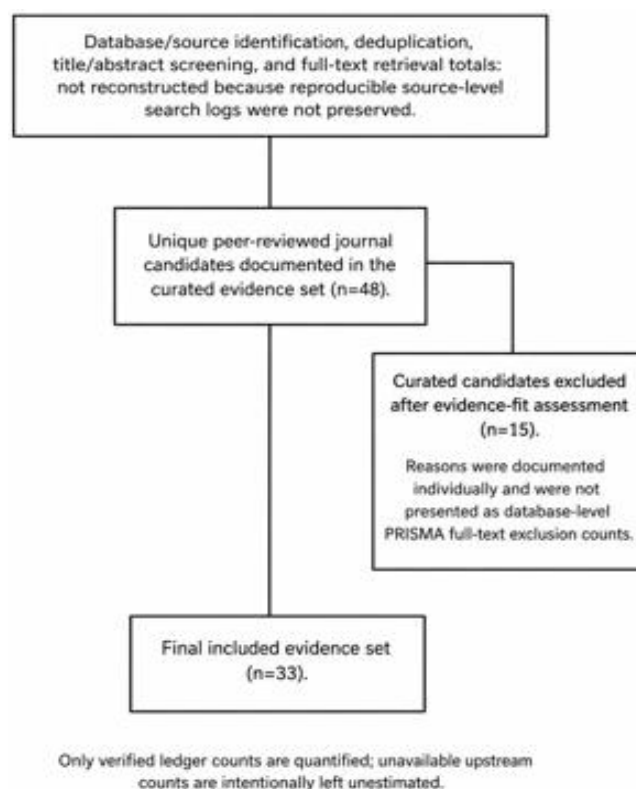


Figure 1. Verified evidence-selection flow for the systematic review

Eligibility criteria

Eligible articles were peer-reviewed journal publications from the prespecified period with a verifiable DOI, relevant journal-quality status, and a direct role in at least one review domain. Eligibility was defined by the prediction question rather than by algorithm alone. Pharmaceutical solvate-propensity modeling was included because it represents a distinct data-driven solid-form task, while remaining different from exact solvate-structure prediction [7]. Multicomponent prediction was also eligible when the endpoint concerned cocrystal or salt formation, but formation likelihood was not treated as evidence of stability, manufacturability, or improved product performance [8].

Studies were excluded when they lacked a direct claim-level role, duplicated another version, fell outside the publication window, or addressed adjacent crystal phenomena without contributing to prediction, validation, uncertainty, or decision interpretation. Case studies were retained only when their evidence resolved a question not better supported by broader or more current work.

Information sources and search strategy

Search concepts combined pharmaceutical solid forms, polymorphism, crystal structure prediction, molecular packing, cocrystals, salts, solvates, crystallization, nucleation, stability, phase transformation, crystal habit, machine learning, deep learning, interatomic potentials,

uncertainty, validation, and developability. Contemporary terminology was expanded to include transfer and foundation-model approaches because molecular-crystal prediction increasingly uses reusable learned representations across downstream tasks [9].

The strategy also included crystal-specific data-curation, representation, sampling, and model-tooling terminology because these steps affect what a model can learn and how benchmark results should be interpreted [10]. Retrieval used cross-publisher scholarly searching, indexed bibliographic records where available, and targeted publisher or journal-platform searches. Searches were iteratively refined around task synonyms and boundary conditions rather than around predetermined table rows or figure metadata. Only records whose DOI, bibliographic metadata, journal eligibility, and claim-level relevance could be verified were retained for evidence mapping.

Study selection

Selection was performed at the level of the scientific question answered by each paper. Studies predicting process behavior were not automatically classified as solid-form prediction studies. For example, machine-learning correlations for primary nucleation kinetics and scale-up address process-level behavior and technology transfer; they do not uniquely determine molecular packing or the final polymorph obtained [11].

Candidate papers were therefore screened for target definition, data source, material class, validation design, and the inference required for inclusion. Preference was given to studies that contributed distinct evidence rather than duplicating a broader point already supported by a stronger source. Recent studies were favored when they offered equivalent or better evidential fit, while older work was retained when it supplied necessary physical, methodological, or task-defining foundations. Duplicate DOI records, version duplicates, non-journal sources, and papers that would require overgeneralization beyond their studied chemical or process domain were excluded.

Data extraction and feature coding

Extraction separated what was predicted from how it was measured or validated. For each included article, coding captured material class, pharmaceutical or analytical setting, data source, prediction target, molecular or crystal representation, algorithm, comparator, train-validation-test design, performance metric, experimental confirmation, uncertainty source, stated limitation, and transferability boundary. When information was absent, it was coded as not reported rather than inferred.

This separation was particularly important for multicomponent systems, where machine-learning prioritization can coexist with experimental cocrystallization evidence in the same workflow [12]. Computational prediction and experimental confirmation were therefore recorded as distinct evidence layers. The extraction also distinguished formation propensity from structural stability, packing prediction from crystallization kinetics, and crystal-level outcomes from downstream formulation consequences. These fields were designed to support task-stratified synthesis rather than a single model leaderboard.

Quality and risk-of-bias appraisal

Prediction-model quality was appraised using an adapted structure informed by PROBAST+AI, while explicitly recognizing that the tool was developed for prediction-model assessment outside molecular-crystal science and is not a validated crystal-specific instrument [13]. The appraisal therefore added domain-specific questions concerning duplicate or near-duplicate structures, molecule- or polymorph-family leakage, random versus grouped or external splits, generated versus experimental structures, uncertain negative labels, class imbalance,

condition omission, reference-theory error, and absence of blind or experimental validation.

Applicability was judged separately from numerical performance. Particular caution was assigned when training and test sets occupied closely related chemical space, when generated crystal candidates were treated as equivalent to experimentally observed forms, or when a model output was interpreted beyond its endpoint. High benchmark accuracy was not considered sufficient evidence of transferable pharmaceutical validity unless the validation design addressed the chemical, structural, and process conditions relevant to the intended decision.

Synthesis approach

Because the included literature addresses noncommensurable endpoints, synthesis was structured without default quantitative pooling. Classification accuracy for a solvate or cocrystal task, ranking quality in crystal structure prediction, force or energy error for an interatomic potential, nucleation prediction, finite-temperature stability, and a measured physicochemical property do not represent the same outcome. Reporting therefore followed principles for transparent synthesis without meta-analysis, while recognizing that SWiM was developed as a general reporting guideline rather than a crystal-model-specific analytical method [14].

Studies were grouped first by prediction target and level of analysis, then by data provenance, representation, validation regime, experimental confirmation, condition dependence, and intended pharmaceutical decision. Methods and review findings were kept distinct in accordance with PRISMA 2020 explanatory guidance [15]. Within each task family, the synthesis emphasized what the reported model directly predicted, what comparison or benchmark supported it, and what inferential step would be required before a downstream developability claim became defensible.

A proposed task-to-decision evidence ladder emerged from this synthesis: computational target definition, model validation, physical or structural plausibility, experimental form confirmation, condition-dependent stability or process behavior, and measured developability consequence. Progress at one level was not assumed to validate the next.

Table 1 summarizes how different prediction contexts require different methodological interpretations and evidence boundaries.

Table 1. Context-by-method crosswalk for interpreting data-driven pharmaceutical solid-form evidence

Prediction context	Typical evidence basis	Appropriate interpretation	Principal boundary
Molecular/form propensity	Molecular or pair descriptors; labeled databases	Screening or prioritization	Propensity does not define exact crystal structure or conditions

Crystal packing/polymorph landscape	Candidate packings; structural descriptors; energy models	Structural plausibility and ranking	Rank does not establish experimental appearance
Crystallization/nucleation	Process variables; kinetic observations	Process-behavior prediction	Process behavior does not uniquely determine final solid form
Atomistic stability/transformation	Crystal structures; energies; forces; trajectories	Thermodynamic or dynamic assessment	Simulated stability does not by itself establish kinetic persistence
Developability-oriented screening	Computational prioritization plus experimental characterization	Decision support when experimentally anchored	Improved solid-state or physicochemical properties do not establish product-level or clinical benefit

Note: The crosswalk is a proposed synthesis derived from the reviewed evidence. Categories describe evidence interpretation rather than validated universal stages.

PRISMA evidence-selection results

The preserved selection record verified a curated set of 48 unique peer-reviewed journal candidates, of which 15 were excluded after evidence-fit assessment and 33 were retained in the final evidence set. These numbers describe the curated candidate set only; upstream database identification, deduplication, title/abstract screening, and full-text retrieval totals were not reconstructed because reproducible source-level logs were unavailable.

The included evidence covered structurally difficult tasks as well as directly pharmaceutical applications. Space-group preference prediction illustrates a target for which molecular information alone may have limited discriminative power and database composition can materially influence apparent predictability [16]. More recent pharmaceutical work extends machine learning to confidence-aware ranking of salts and cocrystals, but a confidence score remains a model output rather than experimental proof that a particular multicomponent form will be realized [17]. Accordingly, the selection results are interpreted by task and evidence level rather than by study count alone.

Characteristics of included studies

The evidence base was methodologically heterogeneous. Some studies used molecular descriptors or learned representations for classification and ranking, whereas others operated at the level of atomic environments, lattice energies, forces, or finite-temperature trajectories. Machine-learned interatomic potentials have enabled molecular-crystal simulations at scales that would be considerably more expensive with direct first-principles calculations, making finite-temperature behavior a practical modeling target [18]. These studies nevertheless inherit dependence on training-domain coverage and on the reference calculations used to generate energies and forces.

A second recent development is the growth of very large generated molecular-crystal datasets. OMC25 exemplifies this shift by coupling large-scale structure generation with benchmark models and explicit split strategies [19]. Such scale is valuable for representation learning and interatomic-potential development, but generated structures are not equivalent to experimentally observed pharmaceutical forms, and random partitions can produce optimistic estimates if related structures occur across sets. Study characterization therefore treated dataset provenance, split logic, experimental status, and level of analysis as core variables rather than supplementary metadata.

Polymorph and solid-form prediction tasks

Polymorph prediction ranged from exploration of candidate landscapes to experimentally anchored solid-form screening. In a drug-specific example, AIMNet2 was used to explore the polymorph landscape of celecoxib, illustrating how neural-network potentials can accelerate evaluation of candidate structures while leaving experimental realization and kinetic accessibility as separate questions [20]. Such studies are most informative when interpreted as landscape or ranking evidence rather than as deterministic forecasts of which form will crystallize.

A stronger evidential connection to development occurs when in silico predictions are tested against experimental solid-form data. Work on a drug-like molecule combined computational screening with experimental validation, demonstrating how modeled hypotheses can be confronted with observed forms instead of being judged only by internal computational criteria [21]. Even this higher tier remains compound- and protocol-dependent because experimental screening may not exhaust all accessible forms.

The recurrent failure modes and their evidence-consistent mitigations are summarized in **Table 2**.

Table 2. Failure modes that weaken pharmaceutical solid-form prediction and evidence-consistent mitigation strategies

Failure mode	Why it matters	Evidence-consistent mitigation
Treating database absence as a true negative	Non-observation may reflect incomplete experimental search	Record label provenance and uncertainty

Random splitting of related structures	Inflates apparent transferability	Use molecule-, family-, or external splits where feasible
Equating computed rank with realized polymorph	Ignores nucleation and kinetic selection	Require experimental or blind validation for realization claims
Ignoring condition dependence	Stability and transformations can change with environment	Evaluate relevant temperature, humidity, solvent, or process conditions
Using model confidence as experimental certainty	Confidence may be miscalibrated or domain-bound	Assess calibration and confirm critical predictions experimentally
Translating solid-state improvement directly to product benefit	Developability has additional formulation and biopharmaceutical determinants	Measure the specific downstream property before extending the claim

Note: The mitigation matrix is a proposed review synthesis; it does not define a validated universal quality standard.

Crystallization and molecular packing models

Molecular packing models and crystallization models solve different problems even when both are described as crystal prediction. Rapid data-driven approaches can use compact topological and physical descriptors to predict structural characteristics of molecular crystals, offering efficient inference over large candidate spaces [22]. Their output, however, concerns packing or structural organization rather than the full thermodynamic and kinetic process by which a particular form nucleates and grows.

Dynamic models add another layer. Machine-learned potentials have been used to investigate single-crystal-to-single-crystal transformations in complex organic molecular systems, extending atomistic simulations to trajectories and system sizes that are difficult to access routinely with direct ab initio dynamics [23]. Such simulations can illuminate possible transformation mechanisms, but the observed pathway remains conditioned by the trained potential, simulation conditions, and the particular crystal system.

The review therefore distinguishes three modeling contexts: static structural inference, process-level crystallization or nucleation prediction, and dynamic atomistic transformation. These contexts can inform one another but are not substitutes. A good packing model may narrow structural hypotheses; a process model may estimate operational behavior; and a dynamic model may

reveal how a transformation can proceed. None alone establishes the complete solid-form outcome under manufacturing conditions.

Stability and solid-state transformation prediction

Stability prediction is especially sensitive to the physical assumptions embedded in the model. Static lattice-energy ranking at or near 0 K may miss vibrational, entropic, and dynamical contributions that become relevant under realistic conditions. Integrated use of crystal structure prediction, lattice dynamics, and molecular dynamics provides a more condition-aware basis for assessing finite-temperature polymorph stability and phase transitions [24]. Even then, calculated thermodynamic preference should not be equated with the rate or probability of transformation.

Environmental conditions add another boundary. Work on crystal-form stability under real-world conditions demonstrates that temperature and humidity can alter stability assessments that would otherwise appear fixed [25]. For pharmaceutical interpretation, this means that “stable form” is incomplete unless the relevant condition range is stated. Practical product stability may also depend on excipients, processing history, particle properties, packaging, and kinetic persistence beyond what a crystal-only model represents.

Table 3 compares the main representations encountered in this evidence base with the assumptions they introduce.

Table 3. Representation–assumption comparison for solid-form, packing, and stability prediction

Representation	Information retained	Key assumption	Important blind spot
Molecular descriptors	Composition and selected molecular features	Molecular properties encode form propensity	Limited direct representation of packing and environment
Learned molecular/crystal embeddings	Data-driven structural regularities	Training data contain transferable patterns	Distribution shift and latent leakage
Candidate crystal structures	Explicit packing geometry	Relevant structures are generated and ranked adequately	Search incompleteness and kinetic accessibility
Interatomic-potential representation	Local atomic environments, energies, forces	Reference calculations are accurate and training coverage is sufficient	Rare interactions and out-of-domain chemistry
Process-variable representation	Operational conditions and measured responses	Historical process relationships transfer	Equipment, scale, and unmeasured-condition dependence
Finite-temperature trajectories/free-energy models	Dynamic and thermodynamic behavior	Sampling and force model capture relevant states	Long-timescale nucleation and formulation effects

Note: The comparison is a proposed synthesis of representation-dependent assumptions across the reviewed evidence.

Developability decision use

A prediction becomes development-relevant only when its endpoint corresponds to a decision that can be tested experimentally. Machine-learning-guided coformer screening for imatinib linked computational prioritization to cocrystal preparation and measured physicochemical properties, providing a stronger decision-use chain than a formation score alone [26]. The evidential step is important: the computational model prioritizes candidates, the experiment establishes whether the solid form can be prepared, and property measurements determine whether the resulting form offers the specific physicochemical advantage claimed. None of these steps independently establishes therapeutic superiority.

Crystal habit illustrates another decision layer. Mechanism-guided, interpretable machine learning has been used to support rational crystal-habit design, connecting process and crystal information to morphology-oriented objectives [27]. Habit can influence filtration, handling, flow, compaction, or downstream processing, but it is not interchangeable with polymorph stability or bulk formulation performance.

The review therefore proposes that developability use should be classified by evidential distance from the model output. Predictions closest to a measurable decision—for example, experimentally confirmed form selection followed by property testing—provide a stronger basis for action than predictions requiring several unvalidated inferential transfers. This does not make them universally transferable; it makes the evidence chain more explicit and testable.

Model validation, generalization, and data gaps

Generalization cannot be inferred from a single random test split. Purpose-built molecular-crystal benchmarks containing experimental and hypothetical polymorphs can expose weaknesses that convenient in-distribution datasets may miss, while global machine-learning potentials directly test whether learned energetic models transfer across structurally and chemically diverse crystals [28, 29]. Together, these studies support a distinction between benchmark accuracy and transferable validity.

Several validation gaps remain recurrent. Molecular or polymorph families may span both training and test partitions; generated structures may dominate evaluation; negative labels may reflect incomplete experimental searches; rare flexible molecules, solvates, salts, and multicomponent systems may be underrepresented; and aggregate errors may conceal failures in the small energy differences that matter for polymorph ranking. External datasets and blind challenges are therefore more informative than repeated tuning against a familiar

benchmark when the intended claim is prospective transfer.

Future validation should also align with decision context. A model intended for pre-experimental solid-form selection should ideally be tested prospectively before new screening results are known. Stability models require condition challenges, not only static structures. Process models require cross-scale or cross-equipment evaluation where transfer is claimed. Uncertainty estimates should be tested for calibration rather than treated as self-validating confidence scores. These requirements are proposed as evidence-alignment principles rather than universal regulatory criteria.

A further gap concerns the distinction between interpolation and decision-level extrapolation. A model may perform well for crystals resembling its training distribution yet fail when a new API introduces unusual flexibility, ionization, hydrogen-bonding motifs, stoichiometry, or solvation. Generalization should therefore be reported along the axis actually tested: new conformers, new molecules, new crystal families, new processing conditions, or new solid-form classes. These axes should not be collapsed into a single claim of robustness. For pharmaceutical use, the most informative test is often the one that withholds the information most likely to be unavailable at the intended point of decision.

Prospective usefulness also requires attention to decision thresholds. A ranking model may be scientifically informative yet operationally unhelpful if several candidates remain experimentally indistinguishable, while a less precise model may still be valuable if it reliably eliminates implausible regions of search space. Evaluation should therefore distinguish statistical performance from decision utility and should report whether uncertainty changes which experiment, form, or process condition would be selected next.

RESULTS AND DISCUSSION

The reviewed evidence shows genuine computational progress, but its pharmaceutical meaning depends on physical scale and validation context. Molecular-crystal polymorphs can differ by small energy margins, and reliable computational description requires attention to many-body intermolecular interactions and the accuracy of the underlying energetic model [30]. This physical difficulty explains why low aggregate prediction error is not automatically sufficient for form ranking: the relevant error is the error relative to the energetic or structural distinction on which the decision depends.

Recent neural-network potentials can substantially reduce the computational burden of crystal structure prediction and stability assessment, including applications evaluated

against demanding targets and blind-test contexts [31]. Their value is therefore not merely speed. They create the possibility of broader landscape exploration and more extensive finite-temperature analysis. Yet efficiency increases the amount of searchable space faster than it guarantees the correctness of every ranking.

The central synthesis proposed here is consequently a task-to-decision evidence hierarchy. Model performance should first be interpreted at its native endpoint, then tested for

structural or physical validity, then connected to experimental realization, condition dependence, and only finally to a stated developability consequence. This hierarchy does not privilege one algorithm family. It instead asks whether each inferential transition is supported by evidence appropriate to that transition. **Figure 2** demonstrates confidence is progressively constrained as prediction moves from computational output toward pharmaceutical decision use.

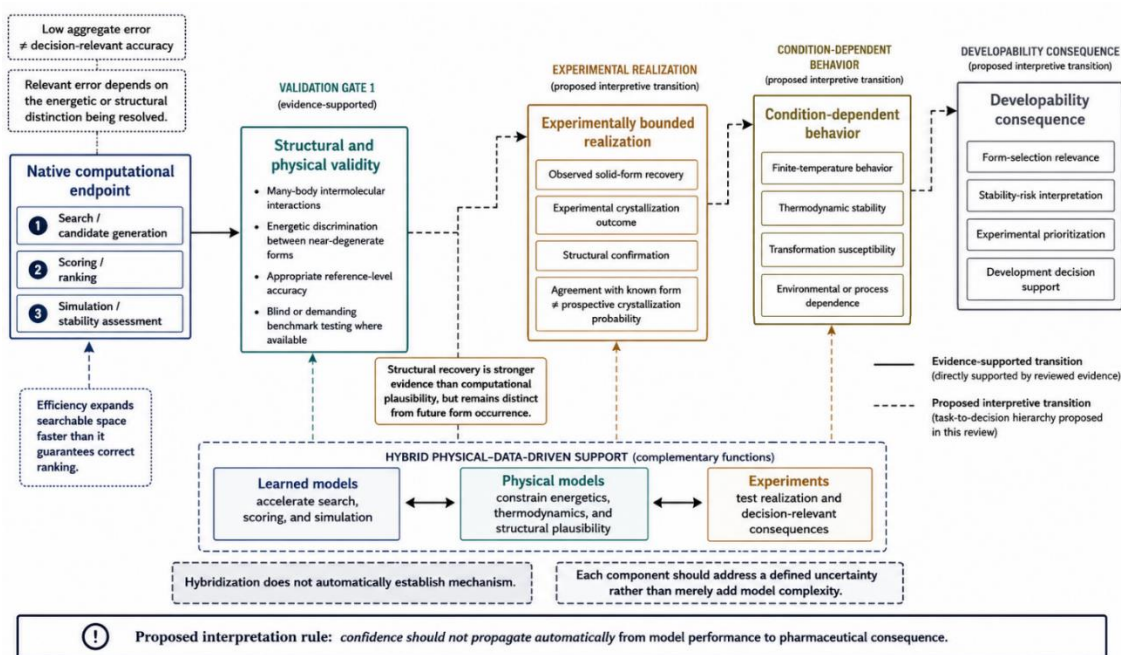


Figure 2. Computational performance and pharmaceutical decision confidence

This hierarchy also clarifies where hybrid physical and data-driven approaches are most defensible. Learned models can accelerate search, scoring, or simulation, whereas thermodynamic models, structural analysis, and experiments can constrain the interpretation of those outputs. The combination should not be described as automatically mechanistic: a physically informed representation may reduce implausible predictions without proving the causal pathway that generated an experimental form. Likewise, agreement between calculation and a known crystal structure is stronger evidence for structural recovery than for future crystallization probability. The practical value of hybridization is therefore conditional—each component should address a defined uncertainty rather than merely adding model complexity.

Review limitations

The review is limited by the rapid evolution of molecular-crystal machine learning and by heterogeneous task definitions, datasets, representations, and evaluation strategies. Geometric deep-learning approaches illustrate how architecture and representation choices can change the

formulation of crystal structure prediction itself, making any cross-study methodological hierarchy time-bounded [32]. Direct numerical comparison is therefore inappropriate when prediction targets or splits differ materially.

Transferability is also task-specific. Hybrid machine-learning approaches for cocrystal stability screening provide evidence within multicomponent crystal spaces, but their performance cannot be assumed for polymorph, solvate, nucleation, or process-control tasks [33]. The evidence base also contains uneven experimental confirmation and limited prospective validation. Finally, the preserved review ledger did not support reconstruction of database-level PRISMA identification and screening counts; those quantities were intentionally left unestimated rather than fabricated.

CONCLUSION

Data-driven pharmaceutical solid-form prediction has developed from isolated property classifiers and conventional crystal-ranking workflows into a

heterogeneous field spanning learned representations, crystal structure prediction, interatomic potentials, crystallization models, finite-temperature stability, transformation analysis, and developability-oriented screening. The principal challenge is no longer simply whether a model predicts something accurately, but whether its target, data, validation, physical assumptions, environmental conditions, and downstream inference are aligned.

This systematic review proposes a task-to-decision evidence hierarchy that keeps computational prediction, physical plausibility, experimental realization, condition-dependent behavior, and pharmaceutical consequence distinct. The framework is intentionally bounded: it does not convert benchmark success into prospective validity or solid-state improvement into product readiness. Progress will depend on representative benchmarks, external and blind testing, prospective experimental confirmation, explicit uncertainty, and validation matched to the actual development decision.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Burcham CL, Doherty MF, Peters BG, Price SL, Salvalaglio M, Reutzel-Edens SM, et al. Pharmaceutical digital design: From chemical structure through crystal polymorph to conceptual crystallization process. *Cryst Growth Des.* 2024;24(13):5417-38. doi:10.1021/acs.cgd.3c01390
- [2] Zhou D, Bier I, Santra B, Jacobson LD, Wu C, Garaizar Suarez A, et al. A robust crystal structure prediction method to support small molecule drug development with large scale validation and blind study. *Nat Commun.* 2025;16(1):2210. doi:10.1038/s41467-025-57479-1
- [3] Lu M, Rao S, Yue H, Han J, Wang J. Recent advances in the application of machine learning to crystal behavior and crystallization process control. *Cryst Growth Des.* 2024;24(12):5374-96. doi:10.1021/acs.cgd.3c01251
- [4] Zhu Z, Zhang Y, Wang Z, Tang W, Wang J, Gong J. Artificial intelligence assisted pharmaceutical crystallization. *Cryst Growth Des.* 2024;24(10):4245-70. doi:10.1021/acs.cgd.3c01408
- [5] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. doi:10.1136/bmj.n71
- [6] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA Statement for reporting literature searches in systematic reviews. *Syst Rev.* 2021;10(1):39. doi:10.1186/s13643-020-01542-z
- [7] Xin D, Gonnella NC, He X, Horspool K. Solvate prediction for pharmaceutical organic molecules with machine learning. *Cryst Growth Des.* 2019;19(3):1903-11. doi:10.1021/acs.cgd.8b01883
- [8] Wang D, Yang Z, Zhu B, Mei X, Luo X. Machine-learning-guided cocrystal prediction based on large data base. *Cryst Growth Des.* 2020;20(10):6610-21. doi:10.1021/acs.cgd.0c00767
- [9] Feng M, Zhao C, Day GM, Evangelopoulos X, Cooper AI. A universal foundation model for transfer learning in molecular crystals. *Chem Sci.* 2025;16(28):12844-59. doi:10.1039/d5sc00677e
- [10] Kilgour M, Tuckerman ME, Rogal J. MXtalTools: A toolkit for machine learning on molecular crystals. *J Chem Inf Model.* 2026;66(7):3402-8. doi:10.1021/acs.jcim.5c02868
- [11] Yerdelen S, Yang Y, Quon JL, Papageorgiou CD, Mitchell C, Houson I, et al. Machine learning-derived correlations for scale-up and technology transfer of primary nucleation kinetics. *Cryst Growth Des.* 2023;23(2):681-93. doi:10.1021/acs.cgd.2c00192
- [12] Birolo R, Bravetti F, Alladio E, Priola E, Bianchini G, Novelli R, et al. Speeding up the cocrystallization process: Machine learning-combined methods for the prediction of multicomponent systems. *Cryst Growth Des.* 2023;23(11):7898-911. doi:10.1021/acs.cgd.3c00696
- [13] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* 2025;388:e082505. doi:10.1136/bmj-2024-082505
- [14] Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ.* 2020;368:l6890. doi:10.1136/bmj.l6890
- [15] Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ.* 2021;372:n160. doi:10.1136/bmj.n160

- [16] Gittins H, Day GM. Can machine learning predict the space group preference of organic molecules? *Cryst Growth Des.* 2026;26(9):3318-26. doi:10.1021/acs.cgd.5c01648
- [17] Ghanavati MA, Khalili B, Alberico D, Rohani S. Machine learning-driven discovery of multicomponent pharmaceutical solid forms via DualNet: Confidence-aware prediction and ranking of salts and cocrystals. *Int J Pharm.* 2025;684:126117. doi:10.1016/j.ijpharm.2025.126117
- [18] Della Pia F, Shi BX, Kapil V, Zen A, Alfè D, Michaelides A. Accurate and efficient machine learning interatomic potentials for finite temperature modelling of molecular crystals. *Chem Sci.* 2025;16(25):11419-33. doi:10.1039/d5sc01325a
- [19] Gharakhanyan V, Barroso-Luque L, Yang Y, Shuaibi M, Michel K, Levine DS, et al. Open Molecular Crystals 2025 (OMC25) dataset and models. *Sci Data.* 2026;13(1):354. doi:10.1038/s41597-026-06628-2
- [20] Zheng P, Abramov YA, Sun CC, Isayev O. Exploring celecoxib polymorph landscape using AIMNet2 machine learning interatomic potential. *Chem Sci.* 2026;17(30):14691-9. doi:10.1039/D5SC09784C
- [21] Muthusamy AR, Chauhan D, Jain AK, Somasundaram MS, Singh A. Navigating solid-form screening using in silico methods validated with experimental data for a drug-like molecule. *Cryst Growth Des.* 2025;25(9):2970-85. doi:10.1021/acs.cgd.4c01740
- [22] Galanakis N, Tuckerman ME. Rapid prediction of molecular crystal structures using simple topological and physical descriptors. *Nat Commun.* 2024;15(1):9757. doi:10.1038/s41467-024-53596-5
- [23] Zhao C, Liu H, Qu DH, Cooper AI, Chen L. A machine learned potential for investigating single crystal to single crystal transformations in complex organic molecular systems. *Chem Sci.* 2025;16(5):2363-72. doi:10.1039/d4sc06467d
- [24] Correa GB, Konstantinopoulos S, Tan BI, Zhang Y, Tavares FW, Adjiman CS, et al. Assessing polymorph stability and phase transitions at finite temperature: Integrating crystal structure prediction, lattice dynamics, and molecular dynamics. *J Chem Theory Comput.* 2025;21(23):12197-213. doi:10.1021/acs.jctc.5c01387
- [25] Firaha D, Liu YM, van de Streek J, Sasikumar K, Dietrich H, Helfferich J, et al. Predicting crystal form stability under real-world conditions. *Nature.* 2023;623(7986):324-8. doi:10.1038/s41586-023-06587-3
- [26] Liang X, Liu S, Li Z, Deng Y, Jiang Y, Yang H. Efficient cocrystal coformer screening based on a machine learning strategy: A case study for the preparation of imatinib cocrystal with enhanced physicochemical properties. *Eur J Pharm Biopharm.* 2024;196:114201. doi:10.1016/j.ejpb.2024.114201
- [27] Liu X, Fu C, Song Y, Wu H, Ding Y, Ma J, et al. Accelerating rational crystal habit design via interpretable mechanism-guided machine learning framework. *Cryst Growth Des.* 2026;26(5):2081-94. doi:10.1021/acs.cgd.5c01644
- [28] Price LS, Paloni M, Salvalaglio M, Price SL. One size fits all? Development of the CPOSS209 data set of experimental and hypothetical polymorphs for testing computational modeling methods. *Cryst Growth Des.* 2025;25(9):3186-209. doi:10.1021/acs.cgd.5c00255
- [29] Žugec I, Geilhufe RM, Lončarić I. Global machine learning potentials for molecular crystals. *J Chem Phys.* 2024;160(15):154106. doi:10.1063/5.0196232
- [30] Hoja J, Ko HY, Neumann MA, Car R, DiStasio RA Jr, Tkatchenko A. Reliable and practical computational description of molecular crystal polymorphs. *Sci Adv.* 2019;5(1):eaau3338. doi:10.1126/sciadv.aau3338
- [31] Nayal KS, O'Connor D, Zubatyuk R, Anstine DM, Yang Y, Tom R, et al. Efficient molecular crystal structure prediction and stability assessment with AIMNet2 neural network potentials. *Cryst Growth Des.* 2025;25(21):9092-106. doi:10.1021/acs.cgd.5c01001
- [32] Kilgour M, Rogal J, Tuckerman ME. Geometric deep learning for molecular crystal structure prediction. *J Chem Theory Comput.* 2023;19(14):4743-56. doi:10.1021/acs.jctc.3c00031
- [33] Wengert S, Csányi G, Reuter K, Margraf JT. A hybrid machine learning approach for structure stability prediction in molecular co-crystal screenings. *J Chem Theory Comput.* 2022;18(7):4586-93. doi:10.1021/acs.jctc.2c00343