Check for updates

View Article Online
View Journal

Soft Matter

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: Y. Uchida, S. Kaji and N. Nakano, *Soft Matter*, 2026, DOI: 10.1039/D6SM00257A.

This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/soft-matter-journal

Please do not adjust margins

View Article Online
DOI: 10.1039/D6SM00257A

ARTICLE

Systematic Error Detection in the Database of Liquid Crystals (LiqCryst) Using Predictive Models

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Yoshiaki Uchida,*[a‡] Shizuo Kaji[b‡] and Naoto Nakano[c‡]

Experimental data often contain anomalies, which can be errors or previously unrecognised knowledge gaps. While errors undermine the reliability of reported findings, unknown gaps can sometimes point to opportunities for discoveries. Machine learning (ML) techniques offer a promising means of identifying such anomalies. In this study, we propose a human-in-the-loop approach that integrates domain expertise and an ML model trained on a comprehensive database of phase transition behaviours of liquid crystalline (LC) materials (LiqCryst 5.2) to scrutinize data integrity. The ML model uncovered multiple anomalies in reported chemical data on LC phase transition behaviours, which were subsequently re-examined by human experts to determine whether they were due to errors. Our results demonstrate that the ML model can effectively detect inconsistencies even within a large-scale database widely regarded as an industry standard. At the same time, anomalies that do not originate from errors may highlight unexplored phenomena and thereby stimulate future discoveries. The proposed methodology for systematically reassessing reported chemical data has the potential to be applied broadly across different materials systems and scientific domains.

Introduction

Quantitative structure-property relationships (QSPR) enable the learning of correlations between molecular structures and physical properties directly from data.[1] Such approaches rely critically on the accuracy of both structural representations and experimental measurements. While crystal structures provide well-defined and reliable molecular information,[2,3] structural identification in soft-matter systems—such as polymers, liquid crystals, and surfactants—is inherently more ambiguous, making these systems more prone to errors in reported data.[4,5] Large-scale databases, including publicly available and commercial ones, have been constructed to support data-driven studies of these materials. However, these datasets inevitably contain errors arising from measurement uncertainties, transcription mistakes, and inconsistent reporting. Such imperfections hinder the development of reliable QSPR models. Therefore, systematic error-detection methods are essential for improving data quality and enabling robust data-driven discovery.[6]

Error detection is a classical yet enduring problem across scientific disciplines.[7] In recent years, machine learning (ML) approaches have proven effective for identifying inconsistencies in large, human-curated datasets, with notable examples including applications to Wikipedia and other knowledge bases.[8] In materials science, ML techniques have advanced rapidly, enabling highly accurate prediction models trained on large-scale datasets.[9,10] These models implicitly encode chemical and physical information present in the data, and can therefore be used to perform meta-analysis to detect anomalous patterns and inconsistencies.[5,6] However, ML techniques alone cannot fully resolve the nature of detected anomalies, as they cannot distinguish between erroneous data and genuinely unknown phenomena. Interpreting such discrepancies requires domain-specific knowledge and contextual understanding of how data are generated and reported. Incorporating human expertise through a human-in-the-loop framework allows these anomalies to be examined in context, providing deeper insight into their origins, as shown in Figure 1. In this work, we

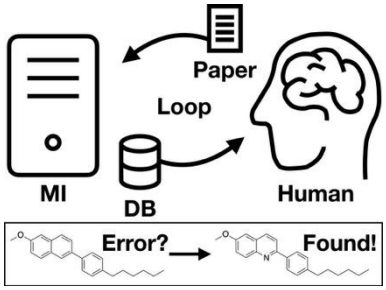

Figure 1. Systematic error detection in a database using predictive models.

[a] Graduate School of Engineering Science, The University of Osaka, 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan. E-mail: y.uchida.es@osaka-u.ac.jp
[b] Graduate School of Science, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto, Japan.
[c] School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano, Nakano-ku, Tokyo, 164-8525, Japan.
‡ Equal contribution.

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Soft Matter Accepted Manuscript

View Article Online
DOI: 10.1039/D6SM00257A

Table 1. Classification Scores and Regression Scores for the Upper and Lower Transition Temperatures of Each LC Phase

Phases	N_V[a]	accuracy [%]	recall	precision	F1 Score	MAE[b] (T_+)	Std[c] (T_+)	MAE[b] (T_-)	Std[c] (T_-)
N, N*	20662	92.4	0.92	0.92	0.92	7.39	11.4	11.0	12.5
SmA	9531	93.6	0.85	0.86	0.85	8.40	12.3	10.1	12.3
SmB	1710	97.9	0.69	0.76	0.72	10.7	12.8	13.3	15.3
SmC, SmC*	6444	96.1	0.87	0.87	0.87	8.00	11.0	8.77	10.7
SmF, SmF*	307	99.7	0.72	0.81	0.76	6.20	8.64	9.11	11.5
SmG	328	99.5	0.63	0.72	0.67	8.54	11.9	10.1	10.9
SmH	52	99.9	0.56	0.62	0.59	13.1	19.0	14.0	13.9
SmI, SmI*	274	99.7	0.75	0.77	0.76	6.13	16.3	10.1	18.0
Col_h	341	99.7	0.78	0.78	0.78	15.7	19.1	17.6	20.0

[a]The numbers of molecules showing the specific phases. [b]Mean absolute error. [c]Standard deviation.

demonstrate such a framework for systematic error detection in materials data, illustrated through several representative case studies.

ML models for liquid crystalline (LC) systems have been developed to predict phase transition behaviour.[10–12] However, previous studies have predominantly relied on carefully curated, small-scale datasets comprising at most a few thousand molecules, where obvious defects are minimised. While such datasets are suitable for model benchmarking, they do not reflect the heterogeneity and noise inherent in real-world data. Recent studies have begun to extend ML-based research on LC materials beyond small, carefully curated datasets, including large-scale analyses and experimentally validated discovery pipelines.[13–15] In this context, our focus is complementary: model-guided detection of database inconsistencies followed by expert re-examination. A key feature of this study is the use of a large-scale database containing over 100,000 substances,[16] which inevitably includes inhomogeneities and various types of anomalies. By training an ML model on this dataset, we construct a predictive model that generalises across a broad range of LC materials while maintaining high accuracy. Leveraging this model, we perform systematic anomaly detection within the database, demonstrating that large-scale, heterogeneous data can be effectively utilised not only for prediction but also for identifying inconsistencies.

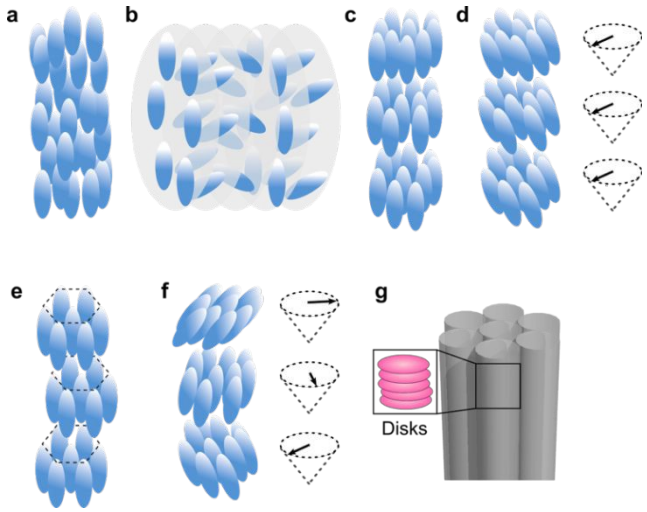

Figure 2. Schematics of representative liquid crystalline (LC) phases. (a) Nematic (N): orientational order without long-range positional order. (b) Chiral nematic (N*): helical structure with the director rotating in space. (c) Smectic A (SmA): layered structure with molecules oriented parallel to the layer normal. (d) Smectic C (SmC): layered structure with the director tilted relative to the layer normal; the in-plane tilt direction is referred to as the C-director. (e) Smectic B (SmB): layered structure with additional short-range hexagonal (bond-orientational) order within the layers. (f) Chiral smectic C (SmC*): helical arrangement of the tilted C-director, combining SmC order with chirality. (g) Hexagonal columnar (Col_h): columns of stacked discotic molecules arranged on a hexagonal lattice.

Methods

We constructed a machine learning (ML) model that takes SMILES representations as input and predicts phase transition sequences. After data curation (described below), the final dataset contained 43,889 molecules. Model performance was evaluated using 5-fold cross-validation to ensure robustness against dataset variability; in each fold, approximately 80% of the data (ca. 35,111 molecules) were used for training and 20% (ca. 8,778 molecules) for testing. All scores were computed from out-of-fold predictions obtained in the 5-fold cross-validation. Thus, each molecule was evaluated by a model that was not trained on that molecule.

The model is based on a graph neural network (GNN), which is well-suited for molecular representation learning due to its ability to capture relational structure beyond handcrafted descriptors. Our implementation is publicly available (https://github.com/shizuo-kaji/LC_QSPR). Specifically, we developed an ML model that predicts both the existence and the temperature ranges of LC phases (Figure 2). The model achieves classification accuracies exceeding 92%, outperforming previously reported approaches for coarse phase classification (typically 80–90%), as summarised in Table 1.[11,12] Furthermore, the prediction error for transition temperatures is within approximately 5%, despite the substantial variability in

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Soft Matter Accepted Manuscript

This journal is © The Royal Society of Chemistry 20xx

View Article Online
DOI: 10.1039/D5SM00257A

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Soft Matter Accepted Manuscript

experimental conditions—such as differences in measurement equipment, sample purity, and heating rates—which can each introduce errors of several degrees Celsius. This level of predictive accuracy enables the reliable identification of anomalous entries in the database.

Predictive model construction

SMILES strings are converted into molecular graphs, where vertices represent atoms and edges represent bonds, with hydrogen atoms omitted to reduce redundancy while preserving essential chemical information. Atom-level features (e.g., hybridisation, atomic mass, formal charge, van der Waals radius) and bond-level features (e.g., bond type, stereochemistry, conjugation) are assigned using RDKit. This choice avoids reliance on predefined molecular descriptors and enables end-to-end learning directly from chemically interpretable primitives. The GNN was implemented with PyTorch Geometric.[17] The model jointly predicts the existence of 13 LC phases (9 achiral and 4 chiral) and their transition temperatures. This multi-task formulation is adopted to exploit correlations between phase occurrence and transition temperatures, thereby improving generalisation compared with training separate models for classification and regression. For the graph convolutional layers, we employ the GATv2 architecture,[18] which allows flexible incorporation of both node and edge attributes through attention mechanisms. This is particularly important for LC systems, where subtle structural variations—such as bond conjugation or stereochemistry—can significantly affect phase behaviour. The network consists of 12 GATv2 layers with interleaved batch normalisation, producing a 256-dimensional latent representation. The network depth was chosen to balance expressivity and overfitting, given the scale and heterogeneity of the dataset.

Two task-specific fully connected heads are used: one for phase classification and one for temperature regression. The model is trained using a combined loss function consisting of focal loss for classification[19] and L1 loss for regression. The focal loss addresses the severe class imbalance caused by rare LC phases, which would otherwise bias the model towards dominant phases under standard cross-entropy. For regression, L1 loss is preferred over L2 loss because the dataset contains non-negligible outliers in reported transition temperatures; L2 loss would excessively penalise such outliers and destabilise training, whereas L1 loss provides greater robustness. Optimisation was performed using the NAdam optimiser,[20] over 300 epochs. This optimiser combines the benefits of adaptive learning rates and Nesterov momentum, leading to stable convergence in practice. Training required approximately 10 hours. Additional implementation details and hyperparameters are provided in the online repository.

Prediction targets

We followed standard QSPR modelling procedures in defining the prediction targets.[21] Specifically, we considered the following LC phases: nematic (N), smectic A (SmA), smectic B (SmB), smectic C (SmC), smectic F (SmF), smectic G (SmG), smectic H (SmH), smectic I (SmI), and hexagonal columnar (Col_h), together with the corresponding chiral phases N*, SmC*, SmF*, and SmI*. The G and H phases are now generally treated as soft-crystal phases rather than smectic LC phases. In the present paper, however, we retain the labels SmG and SmH when referring to these entries to remain consistent with the LiqCryst 5.2 notation. These phases represent the most frequently observed and systematically classified LC phases in the database, and some are schematically illustrated in Fig. 2.

The prediction task was formulated to jointly estimate (i) the existence of each phase and (ii) the corresponding transition temperatures, enabling a consistent description of phase sequences and their thermal behaviour.

Certain phases—namely smectic E (SmE), smectic J (SmJ), and chiral SmJ (SmJ*)—were excluded from the analysis. In the database, the notation for these phases is inconsistent with that used in the original literature, leading to ambiguity in their interpretation. To avoid introducing label noise that could adversely affect model training and evaluation, these phases were not considered as prediction targets.

Within the definition of the prediction targets, chiral and achiral variants were not treated as separate classes. Specifically, the pairs N/N*, SmC/SmC*, SmF/SmF*, and SmI/SmI* were grouped together when defining the phase labels. This choice is motivated by the fact that, within Landau–de Gennes theory of the N–Iso phase transition, the same primary order-parameter framework applies to both N and N* phases; chirality enters mainly through elastic couplings rather than by defining a fundamentally distinct mesophase class at this level of description. A similar argument applies to SmC, SmF, and SmI phases, whose chiral counterparts are often obtained by doping achiral hosts with small amounts of chiral additives. While chirality introduces additional features, such as helix-related signatures in physical properties, phase stability and transition temperatures are typically governed by those of the corresponding achiral host.

To assess whether this treatment affects the results, we performed an additional robustness analysis in which each pair (N/N*, SmC/SmC*, SmF/SmF*, and SmI/SmI*) was reformulated as a three-class problem: achiral, chiral, or absent. The resulting metrics are summarised in Table S1. This analysis confirms that the conclusions remain unchanged: the performance reported in Table 1 is not primarily driven by the merging of chiral and achiral labels. Instead, the dominant limitation arises from class imbalance, particularly for the less common smectic families.

LC Database and data curation

The LiqCryst 5.2 database contains 107,773 substances. Such a large-scale dataset inevitably includes errors, necessitating careful curation for reliable analysis. At the same time, excessive manual intervention can introduce bias, potentially improving performance on the given dataset while reducing generalisability. In this study, our objective is not only to achieve high predictive performance but also to construct a

View Article Online
DOI: 10.1039/D5SM02357A

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

model that captures general characteristics of LC materials from a heterogeneous dataset. This motivates a rule-based and reproducible preprocessing procedure.

Each substance in LiqCryst 5.2 is associated with multiple entries describing its structure and properties. We focused on entries containing molecular structures and phase transition sequences. Molecular structures are provided in several formats (e.g., images, MOL files, line notations, and SMILES). We used SMILES representations because they are well-suited for computational processing. Molecules lacking valid SMILES or containing duplicates were removed; most duplications arose from non-isomeric representations of chiral molecules.

We further defined a reproducible filtering protocol based on phase-sequence criteria and the presence of inorganic or otherwise out-of-scope compounds. Specifically, we imposed a minimum combined count of 12 carbon and nitrogen atoms, approximating the size of two six-membered rings. In addition, all entries containing metallic elements were excluded. These exclusions are intended as a rule-based domain definition and data-quality control, rather than implying that all excluded compounds are intrinsically incapable of exhibiting LC behaviour.

Phase transition sequences are recorded in the "Phases" field, including both lower and upper transition temperatures (e.g., "30 N 48"). We retained only entries with both limits available for the target phases, ensuring consistent comparison with model predictions. Entries with transition temperatures of 0 °C or above 800 °C were excluded as unreliable. We also removed entries containing "0 X 0" or "0 0", which indicate monotropic behaviour.

We included clearing points to isotropic phases ("is"), extrapolated values ("ex"), decomposition temperatures ("dec"), and modified clearing points ("chg"), while excluding ambiguous annotations ("un", "no"). For melting points, only values associated with crystalline phases ("Cr", "Cr1") were used; values associated with polymorphic or glass transitions ("Cr'", "Cr2", "Tg") were excluded.

Special care was taken in handling ambiguous phase labels such as X (unidentified), SmX (uncategorised smectic), and ColX (uncategorised columnar). For example, the entry "Cr 20 X 65 N 72 is" indicates that during the heating process, the substance melts at 20 °C, enters an unknown (unidentified) phase X, transitions to the N phase at 65 °C, and finally transitions to the isotropic phase at 72 °C. For each target phase, entries were excluded if the phase was absent while ambiguous phases of the same class were present, to avoid label uncertainty. Additionally, closely related but distinct phases (e.g., discotic nematic, twist grain boundary phases, antiferroelectric or ferrielectric variants, and non-hexagonal columnar phases) were treated as separate and excluded when necessary to maintain consistency in phase definitions.

After applying these curation steps, we obtained a dataset of 43,889 molecules with valid SMILES, phase assignments, and transition temperatures. The numbers of molecules exhibiting each phase (N_V) were as follows: N and N* (20,662), SmA (9,531), SmB (1,710), SmC and SmC* (6,444), SmF and SmF* (307), SmG (328), SmH (52), SmI and SmI* (274), and Col_h (341).

We emphasise that organometallic compounds were excluded throughout the analysis for all phases; this clarification is particularly relevant for Col_h, where such species are more commonly reported.

Results and discussion

We used the predictive model to reassess the integrity of the database itself. Our strategy is to identify discrepancies between model predictions and reported data—both in phase existence and transition temperatures—and to trace their origins by returning to the original literature. This procedure enables us not only to detect anomalous entries but also to uncover the underlying causes of such inconsistencies. Below, we present representative examples illustrating distinct types of errors.

We first examined compounds exhibiting large discrepancies in transition temperatures. Using the mean and standard deviation (σ) of prediction errors, we identify outliers for the upper transition temperature of nematic phases (T_{N+}). Among compounds reported to exhibit N or N* phases, 13 entries show deviations exceeding the mean error (7.39 °C) by more than 10σ (σ = 11.4 °C), indicating statistically extreme inconsistencies. Examination of the original literature reveals that at least 4 of these entries contain clear errors, while 6 are consistent with the reported data and 3 could not be verified due to unavailable references. Notably, a substantial fraction of these extreme outliers corresponds to genuine errors.

A representative example is shown in Fig. 3a, where the predicted T_{N+} is 58.5 °C, whereas the database reports 388.8 °C, resulting in a discrepancy of −330.3 °C.[22] Interestingly, an alternative reference for the same compound in LiqCryst 5.2 reports T_{N+} = 50 °C,[23] which is consistent with the prediction. Closer inspection reveals that the compound described in the former reference differs from that recorded in the database (Fig. 3b), indicating a transcription error. This example highlights how inconsistencies can arise from incorrect mapping between chemical structures and literature sources, and how such errors can be effectively detected through model-based screening.

We next examined discrepancies in phase existence. Among 18,185 compounds reported to exhibit N phases, 92 are predicted with high confidence (> 0.99) not to exhibit such phases. Detailed inspection of the original papers reveals multiple types of errors. In one case (Fig. 3c), a compound reported as nematic in the database is described in the original paper as exhibiting a smectic A (SmA) phase with identical transition temperatures, indicating a transcription error.[24] In another example (Fig. 3d,e), the molecular structure recorded in the database differs from that in the original publication: a nitrogen atom is replaced by a methine group.[25] This subtle structural discrepancy is sufficient to alter phase behaviour, demonstrating that the ML model captures chemically meaningful distinctions at a level consistent with expert intuition. Similar structural inconsistencies were identified for

Soft Matter Accepted Manuscript

This journal is © The Royal Society of Chemistry 20xx

View Article Online
DOI: 10.1039/D5SM00257A

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

multiple compounds within the same source, suggesting systematic transcription errors.

These observations indicate that high-confidence discrepancies identified by the model are strongly enriched in genuinely erroneous entries, making them effective candidates for targeted data validation.

We also investigate cases with lower prediction confidence, which provide complementary insights. Among 2,477 compounds reported to exhibit N* phases, 162 are predicted not to exhibit N* phases with relatively low confidence (< 0.80). One such example (confidence 0.73) reveals a more complex origin of inconsistency. While the database and the cited paper report identical phase behaviour,[26] further examination of the original conference abstract shows that the compound does not exhibit an N* phase (Fig. 3f).[27] Instead, the reported phase behaviour corresponds to an analogue with a shorter alkyl chain (Fig. 3g), whose data appear to have been incorrectly propagated into the database. This case illustrates that some anomalies arise not from simple transcription errors, but from more intricate chains of misattribution across multiple sources.

Importantly, such multi-step inconsistencies are difficult to identify without combining ML-based anomaly detection with detailed human investigation. This example underscores the complementary roles of statistical detection and domain expertise in understanding the provenance of anomalous data.

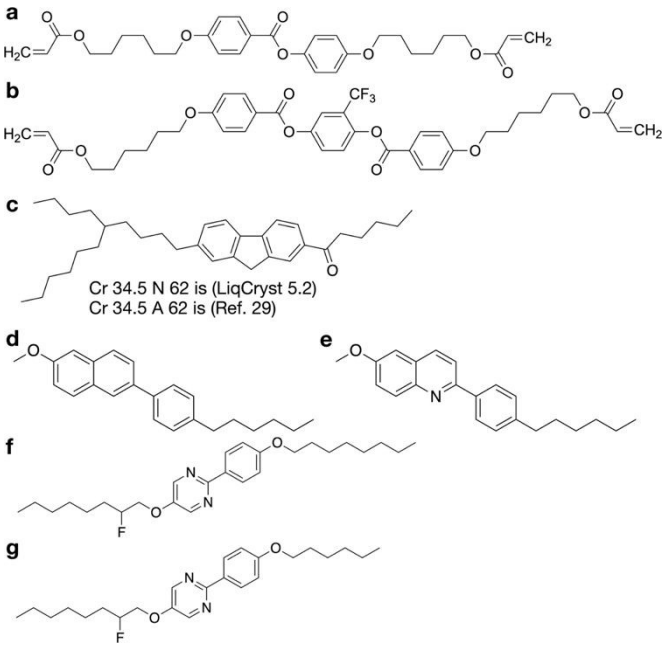

Figure 3. Representative compounds exhibiting significant discrepancies between database entries and machine learning predictions. (a) Compound as recorded in the database and (b) the corresponding structure in the original reference, illustrating a mismatch. (c) Example of incorrect phase assignment: a compound reported as nematic (N) in the database but described as smectic A (SmA) in the literature. (d) Compound as recorded in the database and (e) the corresponding structure in the original reference, highlighting a structural inconsistency. (f) Compound with anomalous transition temperatures in the database and (g) the corresponding structure and data traced to the original source.

Conclusions

Our results demonstrate that multiple types of errors coexist in large-scale LC databases, and that these distinct error modes are reflected in the discrepancies identified by the ML model. Despite such inconsistencies, the majority of entries in LiqCryst 5.2 retain sufficient accuracy to enable the construction of high-performance predictive models. This, in turn, allows the model to serve as an effective tool for systematically detecting erroneous data.

Importantly, not all anomalies necessarily correspond to errors. Some discrepancies may instead reflect unknown or poorly understood phenomena, although identifying such cases lies beyond the scope of the present study. This highlights the dual role of anomaly detection in both data validation and the potential for discovery. Distinguishing between these possibilities, however, remains a challenging task that cannot be resolved solely by predictive models.

The present analysis was conducted on 43,889 of 107,773 LiqCryst entries (40.7%) that satisfied our curation criteria. Within this curated subset, 13 of 20,662 N/N* entries (0.063%) were identified extreme T_{N+} outliers, of which 4 were confirmed as errors, 6 were consistent with the cited data, and 3 could not be verified. For phase-label discrepancies, 92 of 18,185 N-labelled records and 162 of 2,477 N*-labelled records constituted the high-confidence and lower-confidence candidate sets, respectively. These figures describe screened candidate subsets rather than the overall prevalence of database errors.

Our study illustrates that combining ML-based meta-analysis with human expertise provides a practical framework for interrogating large, heterogeneous scientific datasets. Future work will focus on incorporating explainable modelling approaches to better characterise the origins of anomalies and to further bridge the gap between data-driven detection and scientific interpretation.[15]

Author Contributions

Y. U. is the project leader of this work, performed comprehensive study and manuscript drafting and supervises the work. S. K. performed the experiments, comprehensive study, and the writing (review & editing). N. N. developed the methodology and performed the writing (review & editing).

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported in part by the Japan Science and Technology Agency (JST) “Precursory Research for Embryonic

Soft Matter Accepted Manuscript

Science and Technology (PRESTO)" for projects of "Molecular Technology and Creation of New Function" and "Collaborative Mathematics for Real World Issues", Moonshot R&D (Grant Number JPMJMS2021), and JSPS KAKENHI (Grant Number JP20K21226 and JP25H00399). This work is partially supported by FY2020 IMI Joint Usage/Research Program.

Notes and references

(1) Fowles, D. J.; Connaughton, B. J.; Carter, J. W.; Mitchell, J. B. O.; Palmer, D. S. Physics-Based Solubility Prediction for Organic Molecules. *Chem. Rev.* **2025**, *125* (15), 7057–7098. https://doi.org/10.1021/acs.chemrev.4c00855.

(2) Franklin, R. E.; Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **1953**, *171* (4356), 740–741. https://doi.org/10.1038/171740a0.

(3) Beran, G. J. O. Frontiers of Molecular Crystal Structure Prediction for Pharmaceuticals and Functional Organic Materials. *Chem. Sci.* **2023**, *14* (46), 13290–13312. https://doi.org/10.1039/D3SC03903J.

(4) Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011**, *16* (17), 747–750. https://doi.org/10.1016/j.drudis.2011.07.007.

(5) Rodrigues, T. The Good, the Bad, and the Ugly in Chemical and Biological Data for Machine Learning. *Drug Discovery Today: Technologies* **2019**, *32–33*, 3–8. https://doi.org/10.1016/j.ddtec.2020.07.001.

(6) Esaki, T.; Ikeda, K. Data Curation in Cheminformatics: Importance and Implementation. *J Cheminform* **2026**, *18*, 43. https://doi.org/10.1186/s13321-026-01174-w.

(7) Nowogrodzki, J. Cash for Catching Scientific Errors. *Nature* **2024**, *632* (8026), 942–943. https://doi.org/10.1038/d41586-024-02681-2.

(8) Huang, Z.; He, Y. Auto-Detect: Data-Driven Error Detection in Tables. In *Proceedings of the 2018 International Conference on Management of Data*; SIGMOD '18; Association for Computing Machinery: New York, NY, USA, 2018; pp 1377–1392. https://doi.org/10.1145/3183713.3196889.

(9) McDonald, S. M.; Augustine, E. K.; Lanners, Q.; Rudin, C.; Catherine Brinson, L.; Becker, M. L. Applied Machine Learning as a Driver for Polymeric Biomaterials Design. *Nat Commun* **2023**, *14* (1), 4838. https://doi.org/10.1038/s41467-023-40459-8.

(10) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling Deep Learning for Materials Discovery. *Nature* **2023**, *624*, 80–85. https://doi.org/10.1038/s41586-023-06735-9.

(11) Antanasijević, J.; Antanasijević, D.; Pocajt, V.; Trišović, N.; Fodor-Csorba, K. A QSPR Study on the Liquid Crystallinity of Five-Ring Bent-Core Molecules Using Decision Trees, MARS and Artificial Neural Networks. *RSC Adv.* **2016**, *6* (22), 18452–18464. https://doi.org/10.1039/C5RA20775D.

(12) Chen, C.-H.; Tanaka, K.; Kotera, M.; Funatsu, K. Comparison and Improvement of the Predictability and Interpretability with Ensemble Learning Models in QSPR Applications. *J. Cheminform.* **2020**, *12* (1), 19. https://doi.org/10.1186/s13321-020-0417-9.

(13) Piven, A.; Darmoroz, D.; Skorb, E.; Orlova, T. Machine Learning Methods for Liquid Crystal Research: Phases, Textures, Defects and Physical Properties. *Soft Matter* **2024**, *20* (7), 1380–1391. https://doi.org/10.1039/D3SM01634J.

View Article Online DOI: 10.1039/D6SM00257A

(14) Maeda, H.; Wu, S.; Marui, R.; Yoshida, E.; Hatakeyama-Sato, K.; Nabae, Y.; Nakagawa, S.; Ryu, M.; Ishige, R.; Noguchi, Y.; Hayashi, Y.; Ishii, M.; Kuwajima, I.; Jiang, F.; Vu, X. T.; Ingebrandt, S.; Tokita, M.; Morikawa, J.; Yoshida, R.; Hayakawa, T. Discovery of Liquid Crystalline Polymers with High Thermal Conductivity Using Machine Learning. *npj Comput Mater* **2025**, *11* (1), 205. https://doi.org/10.1038/s41524-025-01671-w.

(15) Uchida, Y.; Kaji, S.; Nakano, N. *Chemical-Data-Driven Validation of Physical Theories of Liquid Crystals*. https://doi.org/10.21203/rs.3.rs-1599774/v1.

(16) Vill, V. LiqCryst, 2013.

(17) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*; 2019.

(18) Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? In *International Conference on Learning Representations*; 2022.

(19) Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *42* (2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.

(20) Dozat, T. Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*; 2016.

(21) Roy, K.; Kar, S.; Das, R. N. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, 2015 Edition.; Springer: New York, NY, 2015.

(22) Liang, X.; Cao, H.; Pan, G.; Cui, X.; Li, F.; Niu, G.; Zhang, D.; Yang, Z.; Yang, H.; Zhu, S. Studies on the Electro-optical Properties of Polymer Stabilised Cholesteric Liquid Crystal/Aerosil Particles Composites. *Liquid Crystals* **2009**, *36* (1), 93–100. https://doi.org/10.1080/02678290802680805.

(23) Kurihara, S.; Yoneyama, D.; Nonaka, T. Photochemical Switching Behavior of Liquid-Crystalline Networks: Effect of Molecular Structure of Azobenzene Molecules. *Chem. Mater.* **2001**, *13* (9), 2807–2812. https://doi.org/10.1021/cm0008967.

(24) Malthête, J.; Canceill, J.; Gabard, J.; Jacques, J. Recherches Sur Les Substances Mesomorphes—IX: Smectiques "Fourchus". Organisation Moleculaire Dans Les Phases Smectiques a. *Tetrahedron* **1981**, *37* (16), 2823–2828. https://doi.org/10.1016/S0040-4020(01)92352-X.

(25) Chia, W.-L.; Chang, C. H. Synthesis and Mesomorphic Studies of a Series of Liquid Crystalline 2-(4-Alkylphenyl)-6-Methoxyquinolines. *Mol. Cryst. Liq. Cryst.* **2009**, *506* (1), 47–55. https://doi.org/10.1080/15421400902841452.

(26) Saito, S.; Murashiro, K.; Kikuchi, M.; Demus, D.; Inukai, T.; Neundorf, M.; Diele, S. Ps Inversion in Three Homologous Series of Ferroelectric 5-(2-Fluoro-Alkyloxy)-2-(4-n-Alkylphenyl) Pyrimidines. *Ferroelectrics* **1993**, *147* (1), 367–394. https://doi.org/10.1080/00150199308217206.

(27) Kikuchi, M.; Murashiro, K.; Fukushima, M.; Koyama, S.; Saito, S.; Terashima, K. The Reversal of the Sign of the Spontaneous Polarization of 5-(2-Fluoroalkoxy)-2-(4-Alkyl)Phenylpyrimidine. In *17th Symposium on Liquid Crystals*; 1991; pp 110–111. https://doi.org/10.11538/ekitouyokou.17.0_110.

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Soft Matter Accepted Manuscript

View Article Online
DOI: 10.1039/D6SM00257A

Data availability

The machine learning implementation used to train and evaluate the models—including the graph neural network that takes SMILES representations as input and predicts phase-transition sequences—is publicly available at https://github.com/shizuo-kaji/LC_QSPR. Model performance was assessed using five-fold cross-validation as described in the manuscript. The liquid-crystal phase-transition data analysed in this work were obtained from the LiqCryst database (version 5.2); access is subject to the database provider's licence terms. Additional materials needed to evaluate the findings reported here are available from the corresponding author upon reasonable request.

Open Access Article. Published on 04 May 2026. Downloaded on 5/4/2026 8:32:09 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Soft Matter Accepted Manuscript