



Cite this: *Digital Discovery*, 2026, 5, 3756

Efficient and scalable expansion of property limits via novelty-guided exploration

Satoru Fujii, ^{*a} Ryo Tamura, ^{bcd} Masato Sumita ^{de} and Kei Terayama ^{*adef}

Finding diverse and high-quality candidates in vast design spaces remains an important challenge in scientific research, including molecular design and material discovery. A recent line of work treats observed properties as samples from an unknown distribution and selects candidates that promote more uniform coverage of the property space using Stein discrepancy as the metric. However, these methods do not explicitly account for the information gain of newly acquired data, which can limit exploration of novel regions. In addition, they are not designed for efficient batch selection and may suffer performance degradation when applied in batch settings. Moreover, they incur substantial computational cost due to both large per-iteration overhead and increasing cost as the number of observations grows. To address these limitations, we incorporate predictive uncertainty into the objective function and introduce diversity-aware batch selection strategies to promote informative and diverse candidates. In addition, we propose a sampled Stein novelty estimator and efficient tensor calculations that significantly reduce computational overhead. Performance and runtime evaluations on molecular and materials datasets show that our method enables scalable and sample-efficient expansion of property-space coverage.

Received 19th May 2026

Accepted 2nd July 2026

DOI: 10.1039/d6dd00294c

rsc.li/digitaldiscovery

1 Introduction

Navigating vast design spaces to identify candidates with a wide range of desirable properties remains an important challenge in scientific research, for example, in molecular design and materials discovery, where machine learning and generative models have increasingly been used to guide the search for novel structures and functionalities.^{1–5} In practical applications, it is often desirable not only to optimize specific target properties but also to discover out-of-distribution candidates that cover distinct regions of the property space or Pareto front.^{6,7} In multi-property optimization settings, for example, the objective function is often difficult to define unambiguously. Moreover, even when explicit objectives are available, objective-free exploration can in some cases outperform objective-driven approaches.⁸ Such exploration paradigms have also been used

to obtain representative samples that provide a holistic view of property space,^{9–11} to discover diverse behaviors in robotics,¹² and to efficiently acquire labeled samples that improve predictive models.^{13,14} As illustrated in Fig. 1(a), the overarching goal is to expand coverage beyond densely sampled regions and reveal previously unexplored areas of the property space. Motivated by this background, efficient exploration methods, including active learning^{15,16} and generative models,^{17,18} have been proposed to navigate large design spaces and identify novel materials or molecules with diverse properties.

Since data distributions in property space are often highly biased, naive sampling tends to yield points concentrated in a limited region of the property space. One way to address this issue is to treat observed property values as samples from an unknown distribution and iteratively select candidates that move this empirical distribution toward a uniform distribution over the property space, as illustrated in Fig. 1(b). In particular, methods such as BLOX¹⁹ iteratively select candidates whose predicted properties maximize a quantity called Stein novelty, which measures how much a new point contributes to moving the empirical distribution closer to a uniform distribution, as quantified by the kernelized Stein discrepancy.²⁰ This calculation does not require knowledge of the explicit bounds of the distribution, making it suitable for problems in which the range of the property space is unknown or difficult to estimate.

However, several practical challenges remain. One limitation is that diversity among candidates in the underlying input space (e.g., molecular descriptors or material features) is not explicitly

^aGraduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. E-mail: fujii.sat.rk@yokohama-cu.ac.jp; terayama@yokohama-cu.ac.jp

^bCenter for Basic Research on Materials, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki, 305-0044, Japan

^cGraduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba, 277-8561, Japan

^dRIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan

^eMolNavi LLC, 402 Wizard Building, 1-4-3 Sengen-cho, Nishi-ku, Yokohama, Kanagawa, 220-0072, Japan

^fDepartment of Life Science and Technology, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, 226-8501, Japan

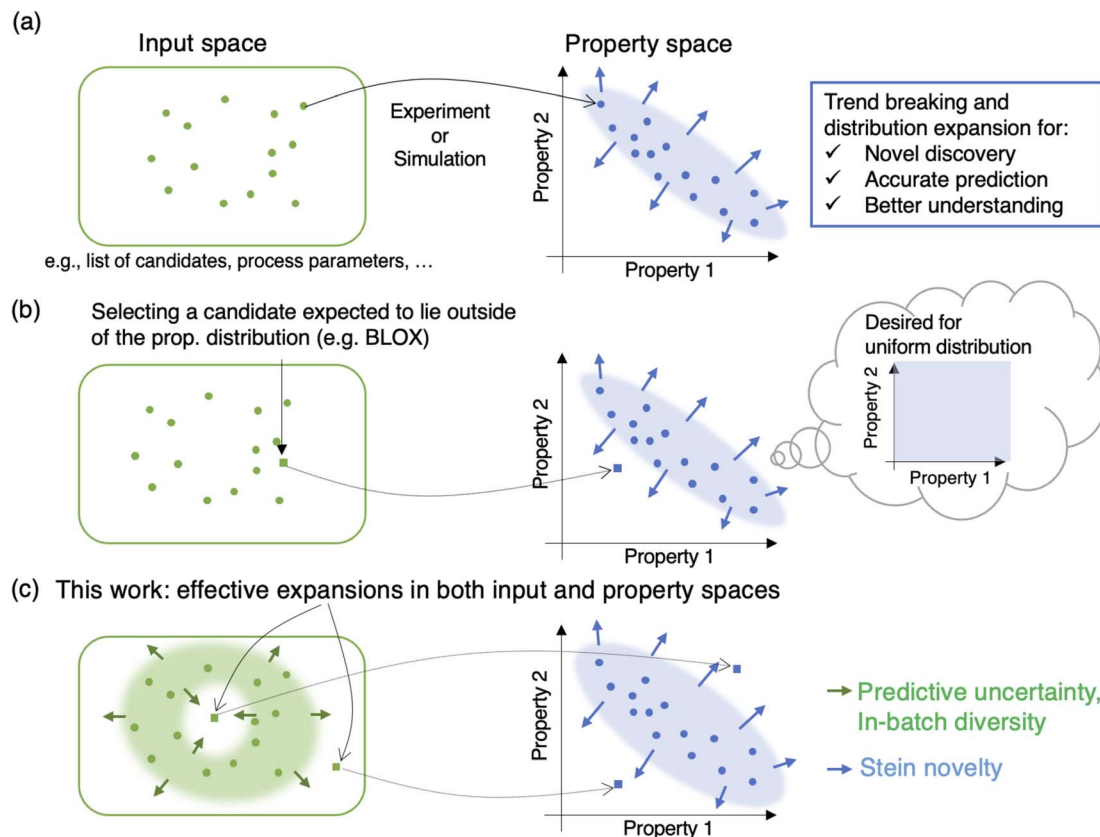


Fig. 1 Overview of property-space coverage expansion. (a) Conceptual illustration of expanding the property space for novel discovery, more efficient learning of accurate predictive models, and better understanding of the data. (b) Illustration of boundary expansion in property space by driving the distribution of observed points toward a uniform distribution. (c) Illustration of improving property-space expansion by combining input-space strategies, including predictive uncertainty and in-batch diversity. Note that although both spaces are shown in two dimensions in this figure, the input space typically has much higher dimensionality than the property space.

considered when selecting new candidates. Incorporating such information has also been studied in prior work,^{13,21} and can be beneficial for two reasons. First, sampling informative points can accelerate the improvement of the predictive model and increase sample efficiency. Second, candidates that are unusual in the input space may correspond to unexplored or extreme regions in the property space. Therefore, considering input-space diversity can help guide exploration toward novel regions of the property space. In this context, we incorporate predictive uncertainty into the objective function to prioritize candidates whose experimental assay is expected to provide high information gain. In addition, we introduce diversity-aware batch proposal strategies that prevent simultaneously proposed candidates from becoming overly similar in the input space. Fig. 1(c) provides an overview of this coverage expansion strategy.

Another challenge is the computational cost of Stein novelty calculations, which grows with the number of observed samples,¹⁹ and limits the applicability of the method to large datasets. Because Stein novelty involves interactions between candidate points and previously observed samples, the computational cost increases linearly as the exploration proceeds. To address this, we propose a sampled Stein novelty

estimator that reduces the asymptotic computational complexity of the algorithm while maintaining the performance of the method. In addition, we implement chunked tensor computations that substantially reduce the runtime of the method.

We evaluate the proposed framework on photo-functional molecules¹⁹ and electronic materials²² in terms of coverage-expansion efficiency and runtime, and show that it enables computationally scalable, sample-efficient exploration. These results highlight the practical utility of the framework for property-space exploration in both sample-limited and large-scale settings. Our framework is freely available as BLOX2 at <https://github.com/ycu-iil/BLOX2>.

2 Methods

2.1 Overview

Our method builds on approaches that select points to expand the empirical property distribution toward a more uniform one, and extends this paradigm in two directions. First, we improve data acquisition by incorporating input-space information through two methods: uncertainty-aware scoring and diversity-aware batch proposal. Second, we accelerate the computation of

the novelty score using two techniques: chunked tensor computation and sampled novelty estimation. An overview of the workflow is illustrated in Fig. 2.

2.2 Stein novelty-based exploration

Given the observed property values, the goal is to select new candidates whose properties expand coverage of the property space.

Following BLOX,¹⁹ we quantify this objective using the Stein discrepancy with respect to a uniform distribution. The contribution of a candidate point to moving the empirical distribution toward the target distribution is measured by Stein novelty.

In the present setting, the property space is bounded by the finiteness of the candidate pool, or the physical and chemical feasibility of the candidates in general. Although its boundary is generally unknown, the Stein discrepancy depends on density gradients rather than explicit boundary specification, so no explicit bound estimation is required during optimization.

The surrogate model predicts a property vector for each candidate, and the Stein novelty score is calculated from the predicted properties. Candidates with higher Stein novelties are considered to contribute more strongly to expanding the coverage of the property space. The details of Stein discrepancy and Stein novelty are provided in SI Text T1 and T2. The Stein novelty score, together with the uncertainty score and the in-batch diversity score when enabled, are calculated for all candidates. After these values are normalized, the acquisition

score is computed as a convex combination of them, and the candidate with the highest score is selected for observation.

2.3 Incorporating predictive uncertainty

Although maximizing Stein novelty promotes coverage expansion in the property space, it does not explicitly account for how newly acquired data improve the predictive model. In active learning settings, especially in Bayesian optimization, uncertainty-aware acquisition is a standard and widely used strategy.^{15,23}

Following this practice, we incorporate predictive uncertainty into the acquisition score. In our experiments, predictive uncertainty was estimated using quantile regression with LightGBM.²⁴ Specifically, the difference between the predicted 0.9 and 0.1 quantiles was used as an uncertainty score.

Uncertainty scores are normalized for each property and then averaged across properties.

2.4 Diversity-aware batch proposal

In many practical settings, each experiment or simulation is costly, so running evaluations in parallel is desirable to reduce wall-clock time. To enable such parallel evaluation, the method must propose multiple candidates at each iteration; therefore, we extend the framework to batch proposal.

Candidates are selected sequentially within each batch. One way to account for candidates already selected within each batch is to use virtual observations. After a candidate is selected, it is treated as a virtual observation with the predicted property value. The acquisition score is then recomputed with respect to

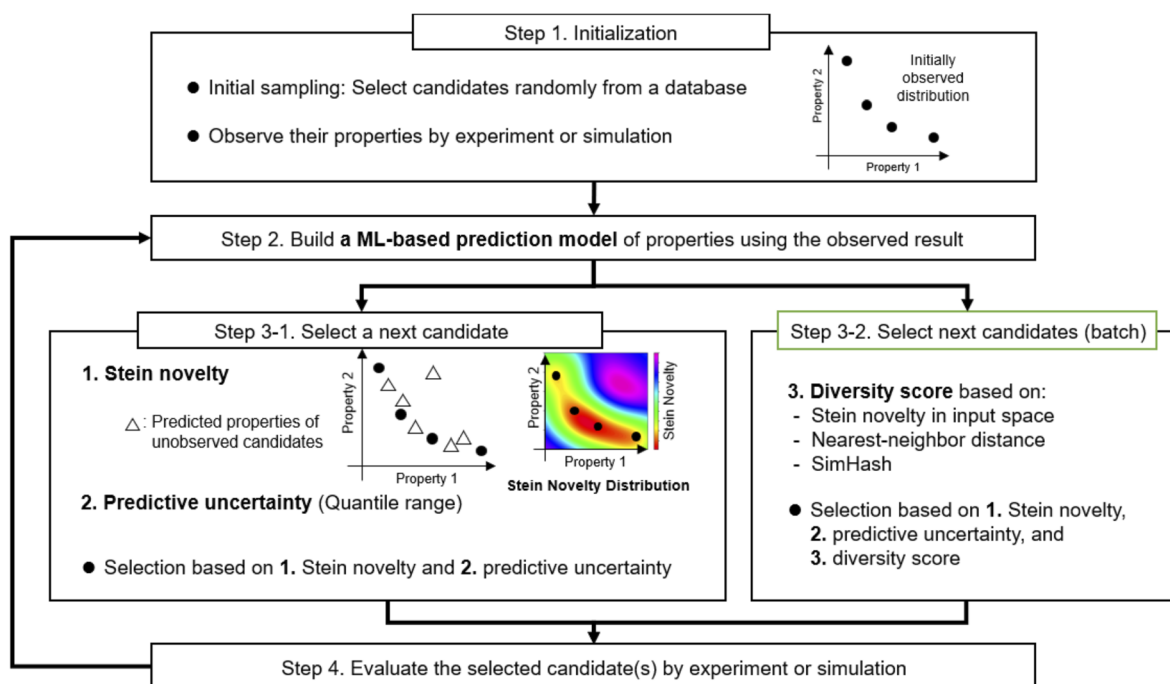


Fig. 2 Workflow of the proposed method. Observed points are used to train the predictor, and new candidates are selected using Stein novelty, calculated in property space from predicted values, together with predictive uncertainty and in-batch diversity control. The selected candidates are then sent to an experimental assay or simulation to obtain new observations for the next cycle.

the augmented set of observations. This procedure is repeated until the given batch size is reached.

However, when simultaneously proposed candidates are overly similar, the acquired batch tends to be redundant and less informative for model improvement. Therefore, rewarding diversity among proposals or penalizing proximity between candidates is widely used in active learning settings.^{25,26} We adopt the same principle here by introducing a diversity score based on proximity in the input space and incorporating it into the acquisition score instead of using virtual observations.

In our experiments, we evaluated three diversity metrics defined on input-space features. Stein-based diversity computes the Stein novelty in the input space with respect to points already selected in the batch and uses it as a diversity term, while Stein novelty in the output (property) space is computed separately for acquisition scoring, as in the other settings. Nearest-neighbor diversity calculates the shortest distance to points already in the batch, similar to greedy sampling,^{13,21} and its multiplicative inverse is used as the score. SimHash-based diversity²⁷ uses random hyperplanes to partition the input space and applies a penalty when a candidate falls into a partition that already contains selected points.

2.5 Accelerating Stein novelty computation

The computational cost of Stein novelty increases as the number of observed samples grows, because each candidate score depends on interactions with the observed set. To maintain scalability on large datasets, we introduce two complementary acceleration techniques that reduce computational time.

2.6 Chunked tensor computation

The dominant cost of Stein novelty computation comes from pairwise interactions between each candidate and the observed samples, which are largely reduced to pairwise distance calculations. In our implementation, these distance calculations are evaluated for chunks of candidate points at once by rewriting the nested loops in the original BLOX implementation as vectorized 3D tensor operations. Although this does not change the asymptotic complexity of the Stein novelty calculation, it substantially reduces computation time in settings where Stein novelty computation becomes a bottleneck. The details of this calculation are provided in SI Text T3.

2.7 Sampled Stein novelty

We further introduce a sampled estimator of Stein novelty. Here, n denotes the number of observed samples, v_p a candidate point, v_i the i -th observed sample, and $u(v_p, v_i)$ the pairwise interaction term in the Stein novelty computation. Instead of using all observed samples, we randomly select a subset $S \subset \{1, \dots, n\}$ of size m for each computation and approximate the full sum by a sum over S . Because a uniformly sampled subset satisfies

$$\mathbb{E}_S \left[\frac{n}{m} \sum_{i \in S} u(v_p, v_i) \right] = \sum_{i=1}^n u(v_p, v_i),$$

the subsampled objective is an unbiased estimator of the full objective, and the multiplicative constant $\frac{n}{m}$ does not affect the minimizer. This approximation reduces the per-iteration computational cost of the Stein novelty estimation from $O(N)$ to $O(1)$ with respect to the number of observations N . A detailed description is provided in SI Text T4.

2.8 Datasets

To evaluate the proposed method in representative multi-objective molecular and materials design settings, we used two benchmark datasets: one for photo-functional molecules and the other for electronic materials. Although the envisioned application is an iterative proposal-to-experimental-assay loop, these fixed datasets allow controlled comparison among methods under the same conditions.

The first dataset comprises 87 268 molecules from the ZINC database,²⁸ with absorption wavelength (nm) and intensity values precomputed by density functional theory (DFT) simulations²⁹ at the B3LYP/6-31G* level. These two properties are directly relevant to the design of photo-functional molecules, where both spectral position and transition intensity are important. As input features, we used Morgan fingerprints with radius 2 and 2048 bits,³⁰ together with RDKit descriptors,³¹ resulting in a total input dimensionality of 2265.

The second dataset consists of 1277 materials obtained from a public database.²² The target properties are band gap (eV) and permittivity, which are fundamental quantities for electronic and dielectric material design. For these materials, we used 132 composition-based physical-property features generated by Magpie.³²

2.9 Metrics

To quantify coverage, we compute the area of the union of ε -neighborhoods around observed samples in the property space. For each dataset, the property space is first normalized using all available true property values, and the resulting area is further divided by the maximum area computed from all available true property values so that coverage takes values in $[0, 1]$. The use of ε -neighborhoods is a commonly adopted formulation.³³ Throughout this paper, we refer to this quantity simply as coverage.

In the main text, we report comparisons for $\varepsilon = 1$ in the normalized property space. Fig. 3 compares the coverage obtained by the proposed method and by random candidate selection, showing that the metric is consistent with the visual extent of the covered region.

When finer local filling behavior is of primary interest, smaller ε values may be more appropriate. We also evaluated the methods with $\varepsilon = 0.1$ and using other evaluation metrics, such as the kernelized Stein discrepancy and convex-hull area, and found broadly similar overall trends across metrics. The corresponding results are provided in SI Fig. S1–S3.

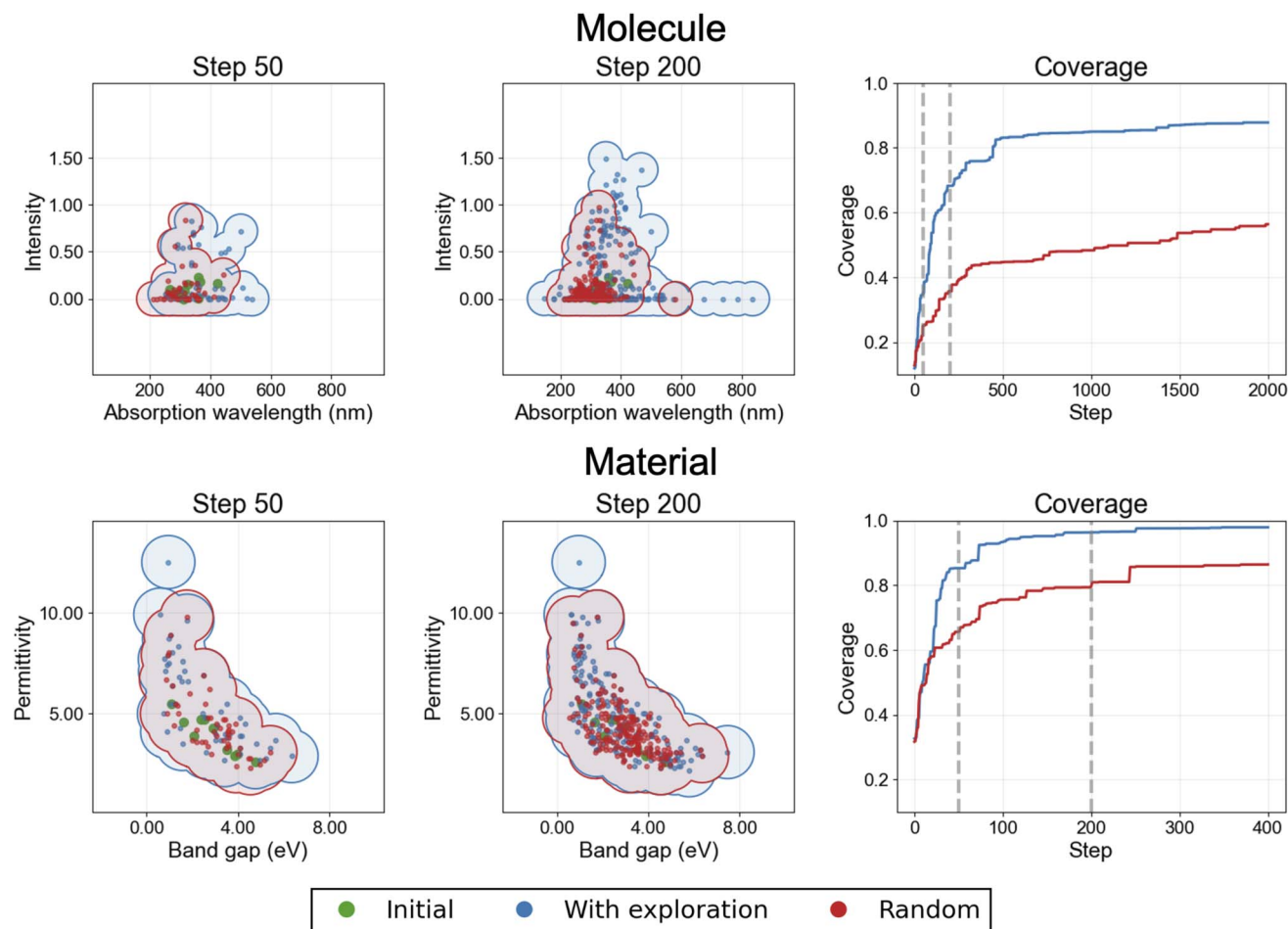


Fig. 3 Illustrative comparison between the proposed method (denoted as “With exploration”) and random candidate selection (denoted as “Random”), demonstrating that the defined coverage metric is consistent with the visual extent of explored regions in the property space. In each row, the left two panels show the observed-property distributions after 50 and 200 proposals, together with the corresponding coverage regions ($\varepsilon = 1$). Here, coverage denotes the area of the union of ε -neighborhoods around observed samples. The right panel shows the trajectory of coverage over steps.

3 Results

3.1 Experimental setup

In the proposed method, a surrogate model is trained on the observed data, and new candidates are iteratively proposed by maximizing Stein novelty computed from the predicted property values. In this section, we evaluate how incorporating predictive uncertainty and in-batch diversity into this selection criterion affects coverage expansion performance, and then assess the computational speedups obtained by the proposed Stein calculation techniques.

All experiments were conducted on the molecular and materials datasets introduced above. For the coverage comparisons, each run was initialized with 10 randomly selected observed points shared across all methods within a seed. We used 5 seeds and 2000 proposals for the molecular dataset, and 20 seeds and 400 proposals for the materials dataset. All samples that were not selected as the initial observed points were used as the candidate pool, and selected candidates were removed from the candidate pool after their

property values were revealed. The reported coverage trajectories are averaged across seeds, and the shaded regions indicate standard deviation. Details of the used hyperparameters are provided in the SI Text T5.

3.2 Effect of predictive uncertainty

We first evaluate the effect of incorporating predictive uncertainty into the acquisition score. In this experiment, uncertainty is defined as the LightGBM quantile spread ($q_{0.9} - q_{0.1}$). The predictive uncertainty-to-Stein novelty ratio was set to 0.75 for both datasets.

Fig. 4(a) compares the coverage trajectories of two strategies for the molecular dataset: selection without model confidence (Stein novelty only) and selection with predictive uncertainty. Fig. 4(b) shows the corresponding results for the materials dataset. The coverage values are averaged across seeds for each dataset.

Across both datasets, incorporating model confidence improved coverage expansion performance. This result

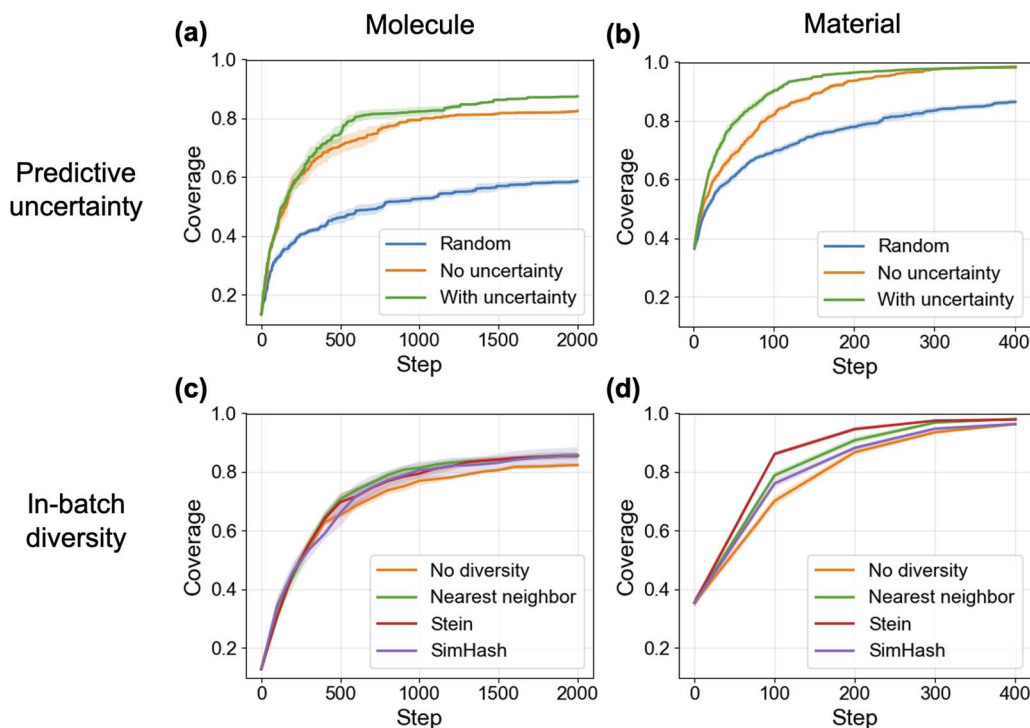


Fig. 4 Results of the predictive-uncertainty and in-batch-diversity experiments on the molecular and materials datasets: (a) predictive uncertainty on the molecular dataset, (b) predictive uncertainty on the materials dataset, (c) in-batch diversity on the molecular dataset, and (d) in-batch diversity on the materials dataset. The curves show the mean coverage trajectories across seeds, and the lightly shaded regions indicate the standard deviation.

indicates that confidence-aware acquisition provides more informative samples and therefore achieves effective exploration of the property space.

3.3 Effect of diversity-aware batch proposal

We next examine batch selection, which is important in practical settings where multiple candidates can be evaluated in parallel. For a fair comparison, predictive uncertainty was disabled in these experiments. We set the diversity-to-Stein novelty ratio to 0.25 for the molecular dataset and 0.50 for the materials dataset, fixed the batch size at 100, and evaluated the following methods:

- Virtual observation baseline without diversity score.
- Stein-based diversity.
- Nearest-neighbor diversity.
- SimHash-based diversity.

Fig. 4(c) shows that incorporating diversity within the batch improves coverage expansion performance compared to raw batch selection for the molecular dataset, and Fig. 4(d) shows the corresponding results for the materials dataset. Because candidate evaluations can often be performed in parallel, batch proposal can further reduce the number of learning iterations required for exploration in practice.

Among the tested methods, Stein-based diversity achieved the best performance. However, because the computational cost of the Stein novelty increases proportionally with the number of target dimensions, it can become expensive in high-

dimensional input spaces. A similar computational-cost concern also applies to the nearest-neighbor distance, although algorithmic refinements may partially mitigate this issue. In contrast, SimHash-based diversity required almost the same computation time as raw batch selection without in-batch diversity, while still improving performance over the raw baseline in both datasets. Considering this practical trade-off, SimHash-based diversity may be a practical default choice. Fig. 5 further shows the effect of Stein-based in-batch diversity on the materials dataset across batch sizes from 1 to 256. These results show that in-batch diversity can substantially suppress the performance degradation associated with increasing batch size, highlighting its importance in large-batch settings.

3.4 Acceleration of Stein novelty computation

Next, we evaluate the effect of the chunked tensor calculation of the Stein novelty. The experiments were conducted on the same machine for fair comparison. Because the materials dataset has both lower input dimensionality and fewer data points, its computation time is trivial and not a practical bottleneck; therefore, the experiments in this section use only the molecular dataset.

Fig. 6(a) shows that our implementation achieves more than a 300-fold speedup of Stein novelty computation compared to the original implementation,¹⁹ enabling efficient evaluation of large candidate sets. Fig. 6(b) presents the same results on a logarithmic scale, showing that the computation time still

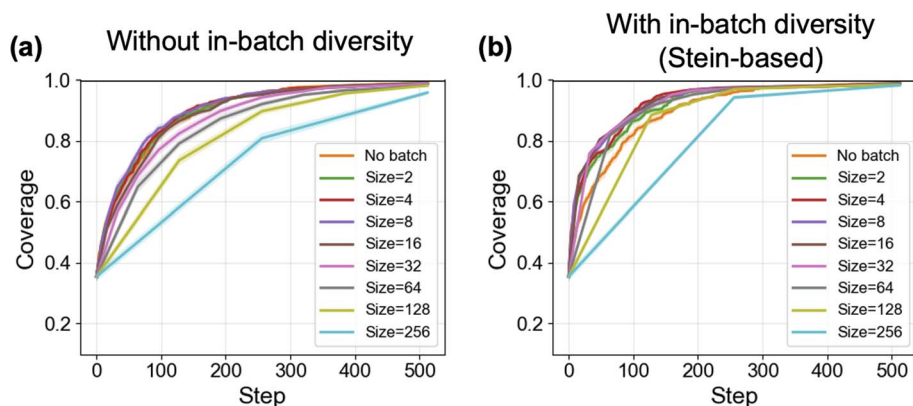


Fig. 5 Effect of Stein-based in-batch diversity as the batch size increases from 1 to 256. Panels (a) and (b) show the mean coverage trajectories across seeds with and without in-batch diversity (Stein-based), respectively, and the lightly shaded regions indicate the standard deviation.

increases approximately in proportion to the number of observed points in both implementations.

Next, we evaluate the sampled Stein novelty estimator. Fig. 6(c) compares the case with sampled Stein novelty (sample size 500) and the case without sampling, showing that sampled Stein novelty makes the computation time constant with respect to the number of observed points. Fig. 6(d) shows that the sampled Stein novelty largely preserves coverage expansion

performance. Even with a sample size of 100, the performance degradation remains small.

4 Discussion

In this study, we extended Stein novelty-based property-space exploration by incorporating input-space information and improving the computational efficiency of novelty calculation. Across molecular and materials datasets, the proposed

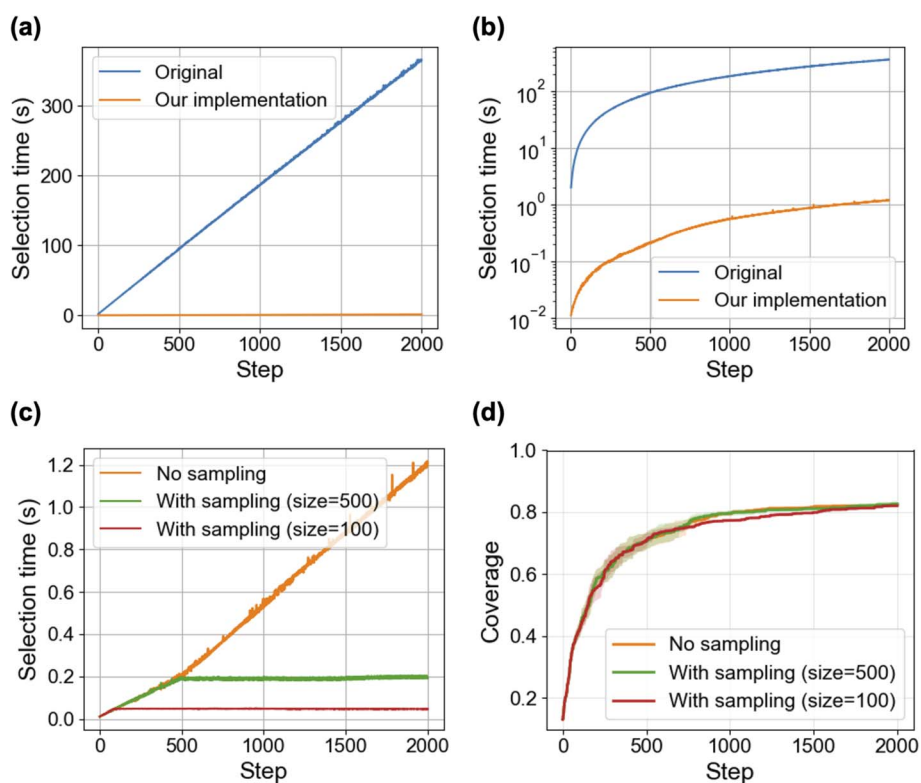


Fig. 6 (a) Comparison of selection time for Stein novelty computation between the original BLOX implementation¹⁹ and our chunked-tensor implementation, measured on the same machine. (b) Shows the same results as (a), but on a logarithmic y-axis. (c) Comparison of selection time for Stein novelty computation with and without sampling (sample size = 500, 100) for the molecular dataset. (d) Comparison of coverage trajectories with and without sampling in Stein novelty computation for the molecular dataset. The lightly shaded regions indicate standard deviation across seeds.

framework achieved sample-efficient coverage expansion together with substantially improved runtime, supporting its practical use in both sample-limited and large-scale settings.

Our method incorporates information from the input space in order to improve the value of the acquired data. In particular, predictive uncertainty is used as a proxy for data value in the acquisition process. An alternative approach is to calculate the Stein novelty directly in the input space. However, its computational cost increases linearly with the dimensionality of the input feature vectors, making it less suitable for high-dimensional datasets. Although we used LightGBM²⁴ in this study due to its favorable balance between accuracy and speed, the framework is not tied to a specific predictor. For example, TabPFN,³⁴ which directly outputs predictive distributions, might further improve exploration quality while also providing uncertainty estimates without separate quantile models, albeit at a higher base computational cost. In our experiments, this strategy produced performance comparable to uncertainty-based selection; details are provided in the SI Fig. S4.

The main practical risk in batch proposal is selecting overly similar candidates. The diversity score introduced in this work mitigates this issue by encouraging exploration within each batch. The computational cost of diversity penalties can be reduced using dimensionality reduction techniques such as principal component analysis (PCA). The squared distance in the original space can be decomposed into the sum of squared distances along the retained and discarded PCA components. Since Stein novelty relies on pairwise distances, its accuracy depends on the fraction of variance captured by the retained components. We also empirically evaluated this approximation, and the results are reported in SI Fig. S5. As predictive performance improves over the course of exploration, the benefit of enforcing in-batch diversity may decrease. Therefore, a scheduling strategy that gradually reduces the weight of the in-batch diversity term as the number of steps increases may further improve overall performance.

The sampled Stein novelty significantly reduces the computational cost of novelty estimation. Although the estimator itself has constant complexity with respect to the number of observations, the overall runtime may still depend on the predictive model used. In many predictive models, retraining the model after each iteration scales at least linearly with the number of observations. Therefore, when scalability is a primary concern, predictive models that support online learning are preferable. Further improvements may be possible by introducing importance sampling strategies³⁵ that focus on informative regions of the property space.

We use a Gaussian kernel in Stein novelty, and its bandwidth is a hyperparameter. In our experiments, the bandwidth is fixed to $\sigma = 1$ for consistency between datasets. In the input space, the feature vectors for all candidate points are known in advance. Therefore, the bandwidth parameters can be estimated using statistics such as the average pairwise distance. In contrast, property values for unobserved candidates are unknown, making such estimation difficult in the property space. Regarding ratios of uncertainty and diversity score, our experiments were conducted on only two datasets and therefore

do not establish a single set of optimal settings applicable to all possible datasets. However, from our experience with the tested settings, an uncertainty-to-Stein novelty ratio of 0.5–0.75 and a diversity-to-Stein novelty ratio of 0.25–0.5 may serve as useful default values.

Our method is applicable to datasets with three or more property dimensions. However, because Stein novelty computation scales linearly with the dimensionality of the property space, the computational cost may become substantial in very high-dimensional settings.

Author contributions

Satoru Fujii: conceptualization (lead); methodology (lead); software (lead); data curation (equal); writing – original draft (lead); writing—review and editing (lead). Ryo Tamura: data curation (equal); supervision (supporting); methodology (supporting); writing—review and editing (supporting). Masato Sumita: data curation (equal); supervision (supporting); methodology (supporting); writing—review and editing (supporting). Kei Terayama: supervision (lead); funding acquisition (lead); methodology (lead); data curation (equal); project administration (lead); conceptualization (lead); writing—review and editing (lead).

Conflicts of interest

There are no conflicts to declare.

Data availability

The source code of BLOX2 and the datasets used in this study are publicly available at <https://github.com/ycu-iil/BLOX2>. The version of the repository corresponding to the experiments reported in this manuscript has been archived on Zenodo and is available at <https://doi.org/10.5281/zenodo.2091113>.

The datasets used in this study are included under the data/directory of the same archived repository. The material dataset was derived from a public oxide dielectric database available at https://github.com/takahashi-akira-36m/oxi_diel_db/tree/master and is provided under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Supplementary information (SI): detailed descriptions of the methods, experimental settings, hyperparameter settings, and additional results. See DOI: <https://doi.org/10.1039/d6dd00294c>.

Acknowledgements

This work was supported by the JST FOREST Program [grant no. JPMJFR232U], JSPS KAKENHI [grant no. 25K01492], and Data Creation and Utilization Type Material Research and Development Project [grant no. JPMXP1122683430].

References

- 1 A. Baranes and P.-Y. Oudeyer, Active learning of inverse models with intrinsically motivated goal exploration in robots, *Robot. Autonom. Syst.*, 2013, **61**(1), 49–73.
- 2 Y. Liu, T. Zhao, W. Ju and S. Shi, Materials discovery and design using machine learning, *J. Materiomics*, 2017, **3**(3), 159–177.
- 3 D. Menon and R. Ranganathan, A generative approach to materials discovery, design, and optimization, *ACS Omega*, 2022, **7**(30), 25958–25973.
- 4 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, Scaling deep learning for materials discovery, *Nature*, 2023, **624**(7990), 80–85.
- 5 D. Nematov and M. Hojamberdiev, Machine learning-driven materials discovery: Unlocking next-generation functional materials—a review, *Comput. Condens. Matter*, 2025, 01139.
- 6 P. Xu, Y. Ma, W. Lu, M. Li, W. Zhao and Z. Dai, Multi-objective optimization in machine learning assisted materials design and discovery, *J. Mater. Inf.*, 2025, **5**, 26.
- 7 Y. Liu, J. Yang, X. Ren, Z. Xinyi, Y. Liu, B. Song, X. Zeng and H. Ishibuchi, Multi-objective molecular design through learning latent pareto set, *Proc. AAAI Conf. Artif. Intell.*, 2025, **39**, 19006–19014.
- 8 J. Lehman and K. O. Stanley, Abandoning objectives: Evolution through the search for novelty alone, *Evol. Comput.*, 2011, **19**(2), 189–223.
- 9 J.-B. Mouret and J. Clune, Illuminating search spaces by mapping elites, *arXiv*, 2015, preprint arXiv:1504.04909, DOI: [10.48550/arXiv.1504.04909](https://doi.org/10.48550/arXiv.1504.04909).
- 10 J. K. Pugh, L. B. Soros and K. O. Stanley, Quality diversity: A new frontier for evolutionary computation, *Front. Robot. AI*, 2016, **3**, 40.
- 11 M. C. Fontaine, J. Togelius, S. Nikolaidis and A. K. Hoover, Covariance matrix adaptation for the rapid illumination of behavior space, in: *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, 2020, pp. 94–102.
- 12 A. Baranes and P.-Y. Oudeyer, Intrinsically motivated goal exploration for active motor learning in robots: A case study, in: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2010, pp. 1766–1773.
- 13 D. Wu, C.-T. Lin and J. Huang, Active learning for regression using greedy sampling, *Inf. Sci.*, 2019, **474**, 90–105.
- 14 H. Liu, B. Yucel, B. Ganapathysubramanian, S. R. Kalidindi, D. Wheeler and O. Wodo, Active learning for regression of structure–property mapping: the importance of sampling and representation, *Digit. Discov.*, 2024, **3**(10), 1997–2009.
- 15 A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, *et al.*, On-the-fly closed-loop materials discovery *via* bayesian active learning, *Nat. Commun.*, 2020, **11**(1), 5966.
- 16 T.-W. Liu, Q. Nguyen, A. B. Dieng and D. A. Gómez-Gualdrón, Diversity-driven, efficient exploration of a mof design space to optimize mof properties, *Chem. Sci.*, 2024, **15**(45), 18903–18919.
- 17 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, Generative models for molecular discovery: Recent advances and challenges, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**(5), 1608.
- 18 M. Cretu, C. Harris, I. Igashov, A. Schneuing, M. Segler, B. Correia, J. Roy, E. Bengio and P. Liò, Synflownet: Design of diverse and novel molecules with synthesis constraints, *arXiv*, 2024, preprint arXiv:2405.01155, DOI: [10.48550/arXiv.2405.01155](https://doi.org/10.48550/arXiv.2405.01155).
- 19 K. Terayama, M. Sumita, R. Tamura, D. T. Payne, M. K. Chahal, S. Ishihara and K. Tsuda, Pushing property limits in materials discovery *via* boundless objective-free exploration, *Chem. Sci.*, 2020, **11**(23), 5959–5968.
- 20 Q. Liu, J. Lee and M. Jordan, A kernelized stein discrepancy for goodness-of-fit tests, in: *International Conference on Machine Learning*, PMLR, 2016, pp. 276–284.
- 21 H. Yu and S. Kim, Passive sampling for regression, in: *2010 IEEE International Conference on Data Mining*, IEEE, 2010, pp. 1151–1156.
- 22 A. Takahashi, Y. Kumagai, J. Miyamoto, Y. Mochizuki and F. Oba, Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations, *Phys. Rev. Mater.*, 2020, **4**(10), 103801.
- 23 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. De Freitas, Taking the human out of the loop: A review of bayesian optimization, *Proc. IEEE*, 2015, **104**(1), 148–175.
- 24 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 3146–3154.
- 25 O. Sener and S. Savarese, Active learning for convolutional neural networks: A core-set approach, *arXiv*, 2017, preprint arXiv:1708.00489, DOI: [10.48550/arXiv.1708.00489](https://doi.org/10.48550/arXiv.1708.00489).
- 26 J. González, Z. Dai, P. Hennig and N. Lawrence, Batch bayesian optimization *via* local penalization, in: *Artificial Intelligence and Statistics*, PMLR, 2016, pp. 648–657.
- 27 M. S. Charikar, Similarity estimation techniques from rounding algorithms, in: *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, 2002, pp. 380–388.
- 28 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, Zinc: a free tool to discover chemistry for biology, *J. Chem. Inf. Model.*, 2012, **52**(7), 1757–1768.
- 29 R. G. Parr, Density functional theory of atoms and molecules, in: *Horizons of Quantum Chemistry*, eds, K. Fukui and B. Pullman, Springer, Dordrecht, 1980, pp. 5–15.
- 30 H. L. Morgan, The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service, *J. Chem. Doc.*, 1965, **5**(2), 107–113.
- 31 Landrum, Greg and others: RDKit: Open-source cheminformatics software, 2019, <https://www.rdkit.org>.
- 32 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Comput. Mater.*, 2016, **2**(1), 16028.

- 33 A. Ghosh and S. K. Das, Coverage and connectivity issues in wireless sensor networks: A survey, *Pervasive Mob. Comput.*, 2008, 4(3), 303–334.
- 34 N. Hollmann, S. Müller, K. Eggensperger and F. Hutter, TabPFN: A transformer that solves small tabular classification problems in a second, in: *The Eleventh International Conference on Learning Representations*, 2023.
- 35 S. T. Tokdar and R. E. Kass, Importance sampling: a review, *Wiley Interdiscip. Rev. Comput. Stat.*, 2010, 2(1), 54–60.