Improved decision making with similarity based machine learning: Applications in chemistry  

Dominik Lemm,1, 2 Guido Falk von Rudorff,1 and O. Anatole von Lilienfeld1, 4, 5, a)  

1) University of Vienna, Faculty of Physics, Kolingasse 14-16, AT-1090 Vienna, Austria  
2) University of Vienna, Vienna Doctoral School in Physics, Boltzmanngasse 5, AT-1090 Vienna, Austria  
3) Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St. George Campus, Toronto, ON, Canada  
4) Machine Learning Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany  

(Dated: 30 November 2023)  

Despite the fundamental progress in autonomous molecular and materials discovery, data scarcity throughout chemical compound space still severely hampers the use of modern ready-made machine learning models as they rely heavily on the paradigm, 'the bigger the data the better'. Presenting similarity based machine learning (SML), we show an approach to select data and train a model on-the-fly for specific queries, enabling decision making in data scarce scenarios in chemistry. By solely relying on query and training data proximity to choose training points, only a fraction of data is necessary to converge to competitive performance. After introducing SML for the harmonic oscillator and the Rosenbrock function, we describe applications to scarce data scenarios in chemistry which include quantum mechanics based molecular design and organic synthesis planning. Finally, we derive a relationship between the intrinsic dimensionality and volume of feature space, governing the overall model accuracy.

I. INTRODUCTION

Useful answers to experimental design1–3 questions are crucial components for successful decision making4–8 under finite budget and time constraints. Conventionally, optimal or near-optimal solutions are proposed by few expert scientists which severely limits humanity’s ‘experimental reach’. As high-dimensional combinatorial solution spaces pose a dramatic bottleneck. Breakthroughs in machine learning and advances in data-availability through high performance computing or high-throughput experimentation are accelerating the pace of scientific discovery and are considered by some to even represent a paradigm-shift.9–11 In particular, autonomous robotic experimentation in the chemical and biological sciences12–17 revolutionizes humanity’s experimental reach through the capacity in which experiments can be executed. The synergy between machine learning, first principles simulations and experimental design facilitates data driven exploration in molecular and materials discovery through an accelerated feedback loop.18 However, rigorous exploration algorithms are in dire need to bypass the combinatorial wall in the solution space. With recent interests in data driven machine learning approaches surging, we have set out to investigate how to exploit modern statistical learning in order to assist with such experimental design exploration challenge19,20 and decision making problems.21 A crucial success factor is the availability of accurate machine learning models, often relying on the increasingly prevalent paradigm ‘the bigger the data the better’ as conveyed through large language model breakthroughs such as OpenAI’s impressive Generative Pre-trained Transformer (GPT-) 4. While GPT-4 is undoubtedly of value through its multimodal capabilities, including novel applications in chemical research22–24 or few shot learning use cases, the overwhelming majority of human endeavours in research are still plagued by intense data scarcity25 often with only hundreds if not only dozens of data points being available, without access to pre-trained all purpose models.

Typically, scarcity is due to high costs imposed by the acquisition of the necessary amount of data. Common examples include long simulation times in quantum chemistry26, cost of compute to perform simulations or money and time to perform laboratory experiments. Consequently, the development of data-efficient machine learning models and training set selection schemes is crucial.27–32 Within the chemical sciences, a concept frequently referenced in drug discovery research33–35 called ‘similar structure-similar property principle’ (SPP) suggests that when the query is known, machine learning models should be trained on instances similar to query, rather than on increasingly many diverse ones (Fig. 1A). Thus, within statistical learning, the training-query similarity emerges as a crucial measure of the predictive power of a model, possibly more relevant than the total quantity of data available.

Here we exploit the concept of SPP in quantum chemical machine learning through introduction of similarity based machine learning (SML) that generates tailored models for arbitrary queries, reaching meaningful prediction accuracy at only a fraction of the total training pool. Ranking prospective data based on similarity to a query, only the nearest neighbours are chosen for training SML models (see Fig. 1B for kernel ridge regression based SML). Note that if not mentioned otherwise, nearest neighbours are selected by proximity rather than all within a fixed distance radius. While conceptually similar to local learning algorithms36 such as k-nearest-neighbors or moving-least-squares, SML models achieve local neigh-
Aquisition of Training Data?

1) Rank by Similarity
2) Train Model on $N$-nearest Neighbours
3) Predict

$y(x_q) = \sum_{i}^{N} \alpha_i K(x_i, x_q)$

FIG. 1. Schematic of similarity based machine learning (SML). (A) As a data acquisition strategy, consider similarity based vs. random selection of costly labels ($y$) corresponding to quantum observables and molecules (represented by wave-functions $\Psi$) from chemical space, resulting in either SML or random machine learning (RML) models, respectively. Due to SML’s exclusive focus on the region most relevant to the query improved data-efficiency is achieved, as evinced by learning curves. (B) Training procedure of SML with kernel ridge regression: Given a vast pool of potentially interesting queries (rectangle), typically only certain queries are of interest. SML selects the corresponding relevant training points ($N_q$) from feature space (banana) as ranked by similarity in representation features (Step No. 1). Consequently, training SML models (Step No. 2) requires costly labels only for the $N_q$ nearest neighbours (shown in circles) which do not yet overlap with any other previously selected query. Resulting query specific SML models yield similar prediction accuracy for less training data (Step No. 3).

In contrast to other data efficient learning schemes such as active learning, SML focuses on few queries at a time rather than training an all purpose model in an iterative fashion. Note that the smaller the number of relevant query instances, the more advantageous SML will be. However, SML can be extended to iterative active learning by using distance or density based training point selection if a multitude of queries is expected. Other data efficient learning algorithms such as transfer or few shot learning require pre-trained models, whereas SML solely relies on training points similar to a query, thereby opening the possibility of query aware few shot learning in domains with limited access to pre-trained models.

For exhaustive screening of chemical compound space, by contrast, conventionally trained random machine learning (RML) models are more appropriate. Note that within this work SML and RML are both based on kernel ridge regression, however, differ in their training data selection and applicability. As such, consider multi-objective compound design in which an RML model is used as a filter to scan vast domains of chemical space, thereby requiring a large and diverse training set in order to answer any query. Once a candidate compound has been identified, i.e. we are aware of the query, SML can subsequently be used to provide highly accurate estimates for properties not covered by the highly general RML model.

After introducing SML in this article, we demonstrate its usefulness for two distinct applications: i) Quantum mechanics based molecular design: Statistical learning of quantum properties holds the promise to accelerate the computational materials design process via fast and accurate machine learning models. Due to the size of chemical compound space (encompassing at least $10^{60}$ compounds) and the computational cost of atomistic simulations (ranging from hours to months), SML can dramatically reduce training data needs if target query compounds are known. ii) Organic synthesis planning: The bottleneck of molecular and materials discovery is amplified in experimental settings, due to large costs and the required manual labour. Additionally, compounds have to be synthesised or purchased before properties can be measured, adding to the cost-time complexity of molecular and materials design. We showcase how SML can be used to efficiently predict relevant properties for target boutique compounds without having make the expenditure necessary to acquire them.
with increasing local nearest neighbour density an decreasing fraction of the total data pool is required to converge to competitive performance.

II. RESULTS

A. Similarity Machine Learning

To assess the effect of the SML approach on learning, we compare learning curves for random and similarity based machine learning models applied to the analytically known harmonic oscillator and the Rosenbrock functions (Fig. 2 A and B). The SML learning curves exhibit dramatically lower offsets and faster convergence when compared to randomly selected ML (RML) models. See methods for more technical details on training and testing.

On average, only very few yet similar training points are necessary for SML to be competitive with the RML model trained on an order of magnitude more training points. In comparison with k-nearest neighbour (k-NN) regression, SML shows on a par performance when the training pool was grid sampled (Fig. 2 B). However, SML outperforms k-NN in randomly sampled training pool settings, suggesting robust performance even in sparsely sampled regions. As mentioned before, since SML is trained on the fly for each query, the use of SML is only advantageous if few queries are of concern. For exhaustive screening and enumeration studies, the conventional RML approach is more appropriate. We note the difference in shape of an SML curve when compared to learning curves reported in the literature. The rapid onset of convergence indicates that only a fraction of available nearest neighbours are necessary to reach the maximum predictive accuracy. Note that in the limit of maximum dataset size, the SML model’s prediction error will always converge towards the corresponding RML prediction error.

B. Dimensionality analysis

We have analysed the SML models throughout this paper using learning curves. The relationship between the predictive accuracy of a model and the number of training points is a fundamental concept in statistical learning. If trained properly, the learning curve as must reflect the inverse relationship between out-of-sample prediction (test) error as a function of number of training points N. Note, in the case of SML based learning curves, N corresponds to the number of nearest-neighbours Nq chosen for training. The logarithmized version of the leading error term is,

\[ \log(\text{Error}) \approx \log(a) - b \log(N), \]

with the slope b being related to the intrinsic dimensionality (d id) and a being related to the target similarity. For the development of data efficient machine learning models it is desirable to understand what decreases log(a) and increases b. Efforts in estimating and analysing the impact of d id have shown an inverse correlation with high model performance, implying that low d id data is generally easier to learn, and that compact representations suffice. In the limit of large N and randomly sampled data, Eq. 1 must become linear for Gaussian Process Regression as well as neural network regressors.

Calculating the distances of all training points in a pool N max to all queries, respectively, the average amount of nearest neighbours Nq within a certain distance radius r can be determined. Fig. 2C confirms the intuition that by increasing the total data pool N max, the local density, i.e. the amount of neighbours within a fixed radius, grows as well. To quantify the relationship between N max and distance radius r, we use the definition of density \( \rho \), i.e. \( \rho = N/V = N/\pi d^2 \), with vol-

![FIG. 2.](image-url)
FIG. 3. Training and analysis of SML models of atomization energies of QM9\textsuperscript{48} compounds using the FCHL\textsuperscript{19,49} representation. (A) Learning curves for RML, SML and k-NN for increasing maximal training set sizes, drawn at random. The inset depicts the error distribution of SML models at varying training set sizes for the largest \( N_{\text{max}} \). Chemical accuracy threshold shown as horizontal line. The nullmodel corresponds to the averaged function value in training set. (B) Learning curves of RML and SML models for predicting 4-methylbicyclo[2.1.0]pentane-2-carbaldehyde (https://tinyurl.com/2p9dtueh) and 3,8-dioxatricyclo[4.2.1.0\textsuperscript{2,5}]nonane (https://tinyurl.com/2p8e4d3h) with either high or low nearest neighbour density, respectively. (C) Display of the first shell (closest ten molecules) and second shell (next closest fifteen) neighbours for the two query compounds.
FIG. 4. Distance based analysis of SML models of learning atomization energies of QM9 using FCHL19 and a Gaussian kernel. \( r_{\text{crit}} \) represents the distance radius at which a certain percentage of the maximally attainable predictive accuracy is reached. (A) Systematic learning of atomization energies using only training instances within the L2-Distance specified on abscissa. (B) Averaged number of nearest neighbours around a query at different total \( N_{\text{max}} \) for increasing L2-distance radius. (C) Percentage of data set size necessary to reach a certain accuracy as determined via \( r_{\text{crit}} \). (D-F) \( r_{\text{crit}}, \) density and intrinsic dimensionality \( d_{\text{id}} \) of the QM9 chemical space as a function of available training points, respectively.

Having introduced the concept of SML and learning curves on mathematical toy functions, we will now apply SML to learning quantum properties of chemical compounds.

C. Quantum mechanics based virtual compound design

To illustrate a meaningful use-case of SML, consider a typical problem of computational molecular and materials design: Organic light-emitting diodes. Consider furthermore that a promising compound would have been identified through sophisticated screening of a very large set of potential candidate compounds. Suppose that the screening imposes all necessary constraints on the candidate compound, except for quantum mechanics based excited states calculations which is not only difficult to model, but is also computationally expensive for larger compounds.\(^{58,59}\) In this scenario, it is difficult to make a rational decision due to the high computational cost and time required to perform the quantum calculation. Furthermore, given the scarcity of literature excited states data throughout chemical compound space, RML models such as state of the art neural network architectures\(^{60-63}\) in quantum chemistry would potentially require larger volumes of data to yield accurate estimates. By contrast, due to its data-efficiency, SML will be able to produce substantially more accurate predictions after training on substantially less data.

Obviously, this molecular design scenario is not restricted to organic electronics. As a numerical illustration, we report the application of SML to learning quantum mechani-
ics based calculated atomization energies (rather than excited states) of organic molecules, as reported in the QM9 dataset, rather than for organic materials candidates. As for the harmonic oscillator and Rosenbrock function discussed above, Fig. 3A displays overwhelming numerical evidence that SML models also exhibit superior data-efficiency when it comes to molecules and their properties. More specifically, for all tested \(N_{\text{max}}\), SML learning curves exhibit a much reduced offset, and converge to the accuracy of the RML learning curve at \(N_{\text{max}}\) for training set sizes that are smaller by orders of magnitude. In contrast, k-NN did not improve beyond the first neighbour. Note also the robustness of SML: As the number of nearest neighbours \(N_{q}\) grows, its predictive variance decreases systematically (Fig. 3A inset). Furthermore, SML performs consistently when applied to quantum properties in chemical compound space as datasets and molecular representations are being varied (for further examples see SI and Fig. S3).

At increasingly large \(N_{\text{max}}\), a shoulder becomes apparent for small training set sizes with less than \(N_{q}=30\). Moreover, the larger \(N_{\text{max}}\), the more data-efficient the SML model (earlier flattening) suggesting that the increase in local neighbour density leads to faster saturation in learning (see II. B for in-depth dimensionality analysis in analogy to the harmonic oscillator and Rosenbrock function). At the largest \(N_{\text{max}}\) possible within the QM9 data-set (~134k molecules for training), SML reaches the coveted chemical accuracy of on average 1 kcal/mol at only 100 training points for any given query. For RML models, this would have solved the ’QM9-IPAM-Challenge’ in quantum chemical machine learning\(^{64,65}\). Note that many applications in quantum mechanics based simulation of molecules and materials are rather concerned with few queries than many. This indicates, that SML might be the superior choice, than learning a general purpose property throughout chemical compound space, such as a universal force field.

To illustrate the effect of the query-specific density on the SML model, we compare two query compounds with respectively low and high density of neighbours in chemical space. More specifically, we consider the caged ether-bridged \(C_{6}H_{10}O_{2}\) compound, 3,8-dioxatricyclo[4.2.1.0\(^5\)8]nonane\(^{66,67}\), and the fused \(C_{7}H_{10}O\) aldehyde, 4-methylbicyclo[2.1.0]pentane-2-carbaldehyde\(^{66,67}\), encoding motifs that are respectively rather rare or frequent within the chemical space spanned by QM9\(^{48}\). SML learning curves (Fig. 3B) confirm the expectation: The higher the density of neighbours the better the learning curve’s off-set and slope in the low-data limit. Note that for query specific training resulting learning curves can also increase with training set size. This is in stark contrast to the statistical learning tenet that averaged prediction errors must decay systematically\(^{42,47}\). Visual inspection of the closest neighbours in terms of FCHL based distances (Fig. 3C) shows barely any relative chemical trends of similarity. Using BoB based distances, however, the visual inspection is more straightforward: Fig. S5 reveals for example that for a high-density query a tautomer in QM9 got selected by SML, resulting in near-perfect predictions.

**FIG. 5.** Example of how to use SML within a virtual custom synthesis scenario. (A) Suppose, the decision to make an investment in a certain desired boutique query compound (exemplary molecular graph and string shown as inset https://tinyurl.com/Syxxzzat) depends on a certain property. Rather than ordering a costly and time-intensive custom synthesis, one can instead simply predict the property by first ordering sufficient more easily available compounds which are similar to the query, and by subsequently measuring their properties in order to train an accurate SML model. The nullmodel corresponds to the averaged function value in training set. (B) Respective learning curves and training data acquisition cost for RML and SML models of an exemplary property (free energies of solvation) using molecules drawn at random from Enamine REAL libraries\(^{46,51}\). Dashed horizontal line indicates chemical accuracy threshold (1 kcal/mol corresponding to experimental uncertainties in thermochemical combustion). (C) Cost vs. time associated with the generation of an SML or RML dataset (~500 US$/compound) in comparison to a ball-park synthesis cost of 2’000 US$/query for different number of queries \(N_{\text{queries}}\) to be expected. The time estimate includes approximate delivery times of 5 days for SML/RML compounds and 1 month for boutique compounds, respectively, as well as an experimental throughput of 25 compounds per day.
D. Dimensionality analysis of QM9

To gain a better understanding of the relation between query and training data proximity, SML models have been trained - similar as reported for the harmonic oscillator and Rosenbrock function - using an increasing distance radius as a measure to select training data. Fig. 4A indicates that as increasing $N_{\text{max}}$ the distance radius required to reach the final accuracy becomes smaller. Concurrently, the average amount of nearest neighbours $\overline{N}_q$ in a smaller radius also increases (Fig. 4B). We define a critical radius $r_{\text{crit}}$ to determine the distance cutoff at which 70, 80, 90, 95% of the maximum accuracy is reached at each $N_{\text{max}}$ (Fig. 4A), respectively. We observe that at increasing $N_{\text{max}}$, $r_{\text{crit}}$ becomes smaller (Fig. 4D). Calculating the average amount of neighbours $\overline{N}_q$ within $r_{\text{crit}}$ (Fig. 4B), a pareto front becomes apparent, confirming that at increasing $N_{\text{max}}$ a fraction of the data is sufficient (Fig. 4C). We apply Eq. 2 on Fig.4B to obtain the density $\rho$ and $d_{id}$. As expected, the density $\rho$ increases with $N_{\text{max}}$ (Fig. 4E). However, in consideration of the size of chemical compound space and ways of systematic enumeration, it is unclear how and if the density will converge and how this will influence the learning of properties in chemical compound space. The estimated $d_{id}$ converges to a value of 2.4 for $N_{\text{max}} > \sim 100$ using a Gaussian kernel (Fig. 4F) and 1.7 using a Laplacian kernel (see SI Fig. S11). While estimates of the $d_{id}$ for popular chemical datasets have not been rigorously reported yet, we note that the $d_{id}$ of a fixed segment of chemical compound space is bound to an upper limit of $4\sqrt{N}$ when using a roto-translational invariant representation. Reasons for potential underestimates can be finite boundaries in terms of representation or even chemical space. Points in high dimensions tend to be localized at the surface of the hypersphere. The effect of edge artifacts can further be amplified through the high dimensionality of molecular representation, which typically are vectors containing thousands of entries. However, we report that even a high dimensional representation such as BoB ($\sim 1'000$ entries for QM9) shows high co-linearity compared to uniform random vectors of the same size (see SI Fig. S12).

E. Organic synthesis planning

A SML use-case in the context of managing synthesis planning within industrial chemistry could be as follows: Consider the problem of acquiring non-trivial molecular properties of one boutique compound through experimental measurements. Synthesis of the compound prior to measurements is also assumed to first require substantial experimental efforts resulting in very long delivery times (also known as 'lead times'). Note that this is not an unusual 'academic' situation as the latter can nowadays be performed commercially through synthesis service companies such as Enamine Ltd. While such services have reached considerable reliability, reporting success rates of 60-80% and higher, challenging boutique target compounds, for example cyclopropylmethyl2-(2-oxo-3,4-dihydro-2H-1,3-benzoxazin-3-yl)acetate, on display in Fig. 5A can impose substantial time delays until the compound has been made, characterized, and shipped, increasing lead times to several weeks or more. Alternatively, SML can be used to make accurate property predictions, rather than measurements, before having to wait out the lead time and measuring the target compound’s properties: First ordering and measuring the most similar compounds with much reduced lead times, for example readily deliverable (3-5 days) building blocks from the Enamine REAL library used for the synthesis of our exemplary boutique compound, would allow for rapidly making the measurements to generate the training data for subsequent SML model generation. The resulting SML models can then be used to accurately infer the property for the original boutique target compound. In comparison to the naive approach of simply waiting for the compound to be delivered, followed by measurements of the properties, this procedure would effectively enable a dramatic cut down of the time to decision while keeping the financial burden for data acquisition at a minimum.

To numerically illustrate the application of SML towards such use cases, we have relied on the aforementioned specific example product (cyclopropylmethyl2-(2-oxo-3,4-dihydro-2H-1,3-benzoxazin-3-yl)acetate), as well as 999 other boutique molecules (see data availability section for a complete list) with lead times of 3 to 4 weeks on average. As a relevant property, we have selected calculated free energy of solvation estimates as a proxy to the measurement. As to be expected from our discussions in the preceding sections, resulting learning curves suggest that on average SML requires a magnitude less data than RML and, similarly, outperforms k-NN (Fig. 5B). Assuming a 500 US$/compound price level on average, estimated purchasing costs for training data-acquisition are shown, also indicating substantial potential for savings, i.e. time is money.

Given that there are multiple queries (up to 1’000), we have also assessed to which degree SML can benefit from synergies due to overlapping nearest neighbours of distinct queries. This possibility would reduce the scaling from the worst case upper bound imposed by RML, rendering SML preferable in terms of time and cost for larger numbers of target queries. To illustrate this scenario, and without loss of generality, we define the target accuracy to correspond to chemical accuracy (1 kcal/mol). Assuming up to 1’000 queries, we have applied SML to select a jointly combined training set containing the closest nearest neighbours of the selected queries. Analysis of the $N_{\text{neigh}}$ required per query to reach 1 kcal/mol (see SI), indicates that less $N_{\text{neigh}}$ are required when considering shared neighbourhood SML models, rather than single query SML. Note that with increasing $N_{\text{queries}}$, also increasingly larger SML training set is required, ultimately converging to the upper bound used within RML models. Thus, variety in the query-set also requires variety training set, whereas few queries are more efficiently predicted via SML.

Additionally, time and cost are constraints to be considered during the design process (Fig. 5C). Albeit the data-acquisition costs of machine learning approaches can be more costly than directly ordering the target compounds (we assume 2’000 US$/target compound price in Fig.5C), SML and RML can become financially advantageous for large sets of
For a modest number of expected queries, SML will be preferable, whereas for many queries RML is the method of choice. In our example, however, the cost per query has not been included as an additional constraint when selecting nearest neighbours. This could lower the total cost even further. In terms of time to decision, SML will always be preferable. Note that the specific variables cost and time strongly depend on each scenario. In conclusion, given additional constraints and a large chemical space to choose from, approximate nearest neighbour selection\textsuperscript{75} is beneficial for reducing cost and time to decision in planning the management of chemical synthesis.

F. Limitations

We want to note that any limitations that apply to an RML model will also apply to SML since both only differ in their training data, but not in their underlying machine learning algorithm, e.g. kernel ridge regression. Since the choice of representation and distance metric influences the choice of nearest neighbours, an improvement in both, e.g. by including more physical knowledge or metric learning\textsuperscript{76}, will result in lower learning curve offsets (see SI Fig. S3). Conversely, low local nearest neighbour density (Fig.3B) will require more data in order to reach optimal accuracy. Additionally, the compute time for an exact nearest neighbour search grows linearly with $O(n)$. However, since the local nearest neighbour density is proportional to the size of search space (see Fig. 4), the $O(\log n)$ scaling of approximate nearest neighbour algorithms, commonly used with millions of data points, becomes more effective. Further strategies could encompass to use efficient low dimensional representations for approximate nearest neighbour searches while using higher quality representations for regression, resulting in a time vs. accuracy trade-off. To make the most out of SML within multi query scenarios, it is recommended to maximize the overlap of shared SML training sets as described in the supplementary information. Methods such as Auto3D\textsuperscript{77}, QOML\textsuperscript{78}, GraphToStructure\textsuperscript{79} or generative models\textsuperscript{80–82} could further facilitate the generation of 3D input structures for quantum chemistry based models to avoid performance loss by using lower levels of theory geometries such as from force fields\textsuperscript{83}. Additionally, the atoms-in-molecule\textsuperscript{84} (AMONS) approach could further aid the applicability of SML by using smaller molecular building blocks to predict significantly larger query compounds.

III. DISCUSSION

We have presented numerical evidence that similarity based machine learning (SML), an approach to select data and train a model on-the-fly for specific queries, can offer substantial improvements in data efficiency for certain use-cases. Applying SML to problems in quantum mechanics and organic synthesis planning, we have found dramatically lower offsets in learning curves\textsuperscript{42}, enabling meaningful decision making at a fraction of the total training set size. Comparisons of SML with other similarity based methods indicate robust performance, even in high dimensional feature spaces. Large training set sizes and associated costs of data acquisition render the use of one-fits-all ML models unfeasible related to the size of chemical compound space. The shape of SML learning curves suggest meaningful deviations from ‘the bigger the better’ paradigm indicating that learning is rather dominated by locality effects, and even saturates upon inclusion of sufficiently many nearest neighbours in training.

After introducing and demonstrating SML for harmonic oscillator and Rosenbrock functions, we have demonstrated its usefulness for quantum mechanics based molecular design where we can reach the coveted 1 kcal/mol chemical accuracy already after just 100 nearest neighbours. This kind of task requires conventional RML models to be trained on thousands of data points. More fundamentally, however, we have derived a relationship between the intrinsic dimensionality and volume of the feature space spanned by the training data which governs the overall model accuracy. In managing the planning of experimental organic synthesis projects, we have shown that SML can bypass time intensive custom synthesis by using readily available compounds similar to the boutique compounds of interest.

We believe that for certain decision problems in chemistry the paradigm of ‘the bigger the data the better’ is not optimal, and that tailored, query specific SML models reach predictive power already for training set sizes that are orders of magnitude smaller than for conventional RML models. More generally, we expect SML to be advantageous for any problem with a) prior knowledge of and access to molecular features in query pool (to easily calculate similarity), b) few queries of interest with labels that are expensive to measure (e.g. long computation times or high costs) c) labels for similar candidates in pool that are less difficult to obtain. Moreover, the implicitly defined domain of applicability of the model provides an indication for the reliability of a machine learning model for unseen queries and inherently limits the usefulness outside of this domain (extrapolation\textsuperscript{85}). Potential applications beyond chemistry could include predictive maintenance tasks in computer aided infrastructure\textsuperscript{86} or aerospace engineering\textsuperscript{87,88}. Future work will address, automated experimental design, deriving ready made models in quantum chemistry or multilevel learning, improving similarity measurements, as well as exploring fast and approximate neighbour searches (cf. turbo similarity fusion)\textsuperscript{76,89–91}.

IV. METHODS

A. Kernel Ridge Regression (KRR)

We rely on kernel ridge regression (KRR)\textsuperscript{92} due to its successes in small data scenarios, however, other machine learning regressors are possible. Kernel based methods belong to the class of supervised machine learning models and are – despite their simplicity – a powerful approach for learning molecules and materials properties. In ridge regression,
a mapping function is learned, relating a representation vector $x$ to a label $y$. The ‘kernel trick’ renders the problem linear through the use of kernel functions that yield the inner product of high dimensional representations $x$. In practice, KRR is formulated as follows:

$$ y(x_q) = \sum_{i} \alpha_i K(x_i, x_q) $$

with $\alpha_i$ being the regression coefficient, $x_i$ being the $i$-th representation vector of the training set, $x_q$ being the query representation and kernel function $K$. The regression coefficients $\alpha$ are calculated through a closed-form solution:

$$ \alpha = (K + \lambda I)^{-1} y_{\text{train}} $$

with the identity matrix $I$ and a regularization coefficient $\lambda$. The latter depends on the anticipated noise in the input and labels and has to be determined through hyperparameter optimization.

### B. Kernel Functions

As a similarity measure, the Laplacian (Eq. 5) and Gaussian (Eq. 6) kernels are the most common choices.

$$ K(x_i, x_q) = \exp \left( -\frac{||x_i - x_q||_1}{\sigma} \right) $$

$$ K(x_i, x_q) = \exp \left( -\frac{||x_i - x_q||_2^2}{2\sigma^2} \right) $$

With $\sigma$ being the kernel width, an additional hyperparameter to optimize. The kernel function for local atomic representations can be rewritten as the sum of pair-wise kernels:

$$ K_{ij} = \sum_{I \in I} \sum_{J \in J} K(x_I, x_J) $$

with $I$ and $J$ being atoms in the training and query, respectively. Kernels for local atomic representations of molecules are often augmented with an additional elemental screening function to compare only atoms of the same nuclear charge:

$$ K(x_I, x_J) = \delta_{Z_I, Z_J} \exp \left( -\frac{||x_I - x_J||_2^2}{2\sigma^2} \right) $$

with Kronecker Delta $\delta$, nuclear charges $Z$ and atomic representations $x$. In order to use local representations as a normalized molecular similarity measure, Eq. 8 can be expressed as the sum of local atomic comparisons:

$$ d_{\text{global}}(x_I, x_J) = \delta_{Z_I, Z_J} \sum_{i \in I, j \in J} x_{i \cdot j} $$

C. Representations

We use standard representations which have been developed, benchmarked and continuously improved over the past years for learning quantum properties. The Coulomb Matrix (CM) contains a 1-body self interaction and a 2-body Coulomb repulsion term and has been one of the first representations developed for machine learning quantum properties. The bag-of-bonds (BoB) representation, the direct successor of the CM, vectorizes the CM terms into fixed size bins, thus enabling the strict comparison between chemically similar environments. The spectrum of London and Axilrod–Teller–Muto (SLATM) representation uses London dispersion contributions as the 2-body and the Axilrod–Teller–Muto potential as the 3-body term and has shown improved performance over BoB for learning quantum properties. The Faber–Christensen–Huang–Lilienfeld (FCHL) representation describes a local atomic environment per atom, respectively, and contains 2 and 3-body terms similar to Behler & Parrinello atomic symmetry functions (ACS) accounting for distance and angular information. We used the optimized FCHL parameters as recommended in Ref. 49. The Smooth Overlap of Atomic Positions (SOAP) representation is, similar to FCHL, a local atomic representation that encodes local geometries through spherical harmonics and radial basis functions. We used $n_{\text{max}} = 3$ and $l_{\text{max}} = 5$ with $\sigma = 0.1$ and a cutoff of 3.5 Å as SOAP settings. For the learning of the free energy of solvation, we used FCHL and the free-energy-machine-learning (FML) scheme.

D. Training

We optimize the hyperparameters $\sigma$ and $\lambda$ through cross-validation on the largest training set size. Both parameters are held constant for all other training set sizes. To train an SML model, the pair-wise distances of a query to all available datapoints defined by $N_{\text{max}}$ are calculated. An overview of training/query set sizes and distance metrics for each representation and dataset can be seen in Tab. S1 in the SI. After sorting all points by distance, only the $V_q$ closest neighbours are considered for training. SML learning curve are obtained by averaging across all single query runs. Similarly, k-NN baselines are calculated by averaging the labels of the closest $V_q$ neighbours weighted by the distance to the query, respectively. RML models are trained choosing $N$ data points at random. The regression coefficients are obtained through Eq. 4 and the query label predicted via Eq. 3 (Fig. 1 B). Note, RML models are only trained once to predict the labels of all queries, whereas SML and k-NN models are trained on-the-fly for each query, respectively.
E. Data

To assess SML, several quantum based datasets have been considered. The QM7 dataset totals 7,165 organic molecules up to seven heavy atoms (CONS)\(^{94,99}\). Atomization energies and structures were calculated using the Perdew-Burke-Ernzerhof hybrid functional (PBE0) level of quantum chemistry\(^{94,99}\). The QM9 dataset\(^{48}\) is a hallmark benchmark in quantum machine learning. Containing 133,885 small organic molecules up to nine heavy atoms (CONF), properties were calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry\(^{48}\). Additionally, a subset of 6,095 C\(_2\)O\(_2\)H\(_10\) constitutional isomers has been used as a toy case.

The Enamine REAL database is the largest enumerated database of synthetically available compounds containing over 4.5 billion molecules\(^{80,51}\). We have chosen 1,000 random query compounds of the REAL library containing 6-21 heavy atoms. For training, we used the Enamine REAL building blocks, which encompass over 128,856 compounds that are used to synthesize the Enamine REAL dataset. We applied a SMILES\(^{100}\) standardization protocol and filtered all compounds containing B, Si or Sn elements, yielding a total number of 124,440 compounds. 3D coordinates and conformers for all compounds were generated with ETKDG\(^{101}\). A maximum of 25 conformers were further relaxed using GFN2-\(\text{xtB}\)\(^{102}\) and corresponding Boltzmann weights obtained through energy rankings. As a training label, free energies of solvation were computed using the Reaction-Mechanism-Generator\(^{73,74}\).

V. ACKNOWLEDGEMENT

Icons used in Figures 1, 5 and S16. were created by Freepik, photo3idea_studio, Victoruler, Maxim Basinski Premium and Monkik of flaticon.com. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 772834). This research is part of the University of Toronto’s Acceleration Consortium, which receives funding from the Canada First Research Excellence Fund (CFREF). Part of this research was performed while GFvR was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1925919).

VI. AUTHOR CONTRIBUTIONS

DL wrote new software used in the work, produced all figures, performed the literature search, compiled the references. DL and GvR acquired the data. DL, GvR and OAvL conceived and planned the project, analyzed and interpreted the results, and wrote the manuscript.

VII. DATA AND CODE AVAILABILITY

The QM9 dataset is available at https://dx.doi.org/10.6084/m9.figshare.c.978904.v5. The QM7 dataset is available at http://quantum-machine.org/datasets/. The Enamine REAL database and building blocks can be accessed at https://enamine.net/compound-collections/real-compounds/real-database. The code to produce SML learning curves is available at https://doi.org/10.5281/zenodo.7870923.

VIII. CONFLICT OF INTEREST

The authors have no conflict of interest.

IX. REFERENCES

1. R. A. Fisher, “Design of experiments,” British Medical Journal, vol. 1, no. 3923, p. 554, 1936.
2. K. Chaloner and I. Verdinelli, “Bayesian experimental design: A review,” Statistical science, pp. 273–304, 1995.
3. F. Pukelsheim, Optimal design of experiments. SIAM, 2006.
4. W. Edwards, “The theory of decision making.,” Psychological bulletin, vol. 51, no. 4, p. 380, 1954.
5. J. W. Pratt, H. Raiffa, and R. Schlaifer, Introduction to statistical decision theory. MIT press, 1955.
6. J. O. Berger, Statistical decision theory and Bayesian analysis. Springer Science & Business Media, 2013.
7. D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, “The statistical complexity of interactive decision making,” arXiv preprint arXiv:2112.13487, 2021.
8. M. Strommershäuser, L. T. Maloney, and M. S. Landy, “Decision making, movement planning and statistical decision theory,” Trends in Cognitive Sciences, vol. 12, pp. 291–297, Aug. 2008.
9. T. Hey, S. Tansley, and K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, October 2009.
10. O. A. von Lilienfeld, “Quantum Machine Learning in Chemical Compound Space,” Angewandte Chemie International Edition, vol. 57, pp. 4164–4169, Apr. 2018.
11. O. A. von Lilienfeld, “Introducing machine learning: Science and technology,” Machine Learning: Science and Technology, vol. 1, p. 010201, Feb. 2020.
12. R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, “Functional genomic hypothesis generation and experimentation by a robot scientist,” Nature, vol. 427, pp. 247–252, Jan. 2004.
13. R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, “The automation of science,” Science, vol. 324, pp. 85–89, Apr. 2009.
14. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, and A. I. Cooper, “A mobile robotic chemist,” Nature, vol. 583, pp. 237–241, July 2020.
15. F. Häse, L. M. Roch, and A. Aspuru-Guzik, “Next-generation experimentation with self-driving laboratories,” Trends in Chemistry, vol. 1, pp. 282–291, June 2019.
16. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, “Controlling an organic synthesis robot with machine learning to search for new reactivity,” Nature, vol. 559, pp. 377–381, July 2018.
of Machine Learning Research, pp. 9323–9332, PMLR, 18–24 Jul 2021.

61K. Atz, C. Isert, M. N. A. Böcker, J. Jiménez-Luna, and G. Schneider, “A-quantum machine-learning for medicinal chemistry,” Physical Chemistry Chemical Physics, vol. 24, no. 18, pp. 10775–10783, 2022.

62Y.-L. Liao, B. Wood, A. Das, and T. Smidt, “Equiformer2: Improved equivariant transformer for scaling to higher-degree representations,” arXiv, 2023.

63P. Thölke and G. D. Fabritiius, “Equivariant transformers for neural network based molecular potentials,” in International Conference on Learning Representations, 2022.

64O. A. von Lilienfeld and K. Burke, “Retrospective on a decade of machine learning for chemical discovery,” Nature Communications, vol. 11, Sept. 2020.

65O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, “Exploring chemical compound space with quantum-based machine learning,” Nature Reviews Chemistry, vol. 4, pp. 347–358, July 2020. Number: 7 Publisher: Nature Publishing Group.

66K. Rajan, A. Zielensy, and C. Steinbeck, “STOUT: SMILES to IUPAC names using neural machine translation,” Journal of Cheminformatics, vol. 13, apr 2021.

67D. Lemm, G. F. von Rudorff, and A. von Lilienfeld, “Leruli.com, online molecular property predictions in real time and for free.” www.leruli.com, 2021.

68K. Molga, S. Szymkú, and B. A. Gryanowski, “Chemist ex machina: Advanced synthesis planning by computers,” Accounts of Chemical Research, vol. 54, pp. 1094–1106, Jan. 2021.

69C. W. Coley, W. H. Green, and K. F. Jensen, “Machine learning in computer-aided synthesis planning,” Accounts of Chemical Research, vol. 51, pp. 1281–1289, May 2018.

70I. Levin, M. Liu, C. A. Voigt, and C. W. Coley, “Merging enzymatic and synthetic chemistry with computational synthesis planning,” Nature Communications, vol. 13, Dec. 2022.

71B. Mikulak-Kucznik, P. Golkiewska, A. A. Bayly, O. Popik, T. Kuczkin, S. Szymkę, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheitl, K. Molga, J. Mylnarski, M. Mrksich, and B. A. Gryanowski, “Computational planning of the synthesis of complex natural products,” Nature, vol. 588, pp. 83–88, Oct. 2020.

72J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, “Principles of early drug discovery,” British Journal of Pharmacology, vol. 162, pp. 1239–1249, Feb. 2011.

73Y. Chung, R. J. Gillis, and W. H. Green, “Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data,” AIChE Journal, vol. 66, Apr. 2020.

74Y. Chung, F. H. Vermeire, H. Wu, P. J. Walker, M. H. Abraham, and W. H. Green, “Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy,” Journal of Chemical Information and Modeling, vol. 62, pp. 433–446, Jan. 2022.

75J. Reis and D. Lowe, “Shape indexing using approximate nearest-neighbour search in high-dimensional spaces,” in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1000–1006, 1997.

76R. Fabregat, P. van Gerwen, M. Haeberle, F. Eisenbrand, and C. Corminboeuf, “Metric learning for kernel ridge regression: assessment of molecular similarity,” Machine Learning: Science and Technology, vol. 3, p. 035013, Sept. 2022.

77Z. Liu, T. Zabatiuk, A. Roithberg, and O. Isayev, “Auto3d: Automatic generation of the low-energy 3d structures with ANI neural network potentials,” Journal of Chemical Information and Modeling, vol. 62, pp. 5373–5382, Sept. 2022.

78S. Heinen, G. F. von Rudorff, and O. A. von Lilienfeld, “Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning,” The Journal of Chemical Physics, vol. 157, p. 221102, 12 2022.

79D. Lemm, G. F. von Rudorff, and O. A. von Lilienfeld, “Machine learning based energy-free structure predictions of molecules, transition states, and solids,” Nature Communications, vol. 12, July 2021.

80E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, “Equivariant diffusion for molecule generation in 3D,” in Proceedings of the 39th International Conference on Machine Learning (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of Proceedings of Machine Learning Research, pp. 8867–8887, PMLR, 17–23 Jul 2022.

81M. Xu, A. S. Powers, R. O. Dror, S. Ermon, and J. Leskovec, “Geometric latent diffusion models for 3D molecule generation,” in Proceedings of the 40th International Conference on Machine Learning (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of Proceedings of Machine Learning Research, pp. 38592–38610, PMLR, 23–29 Jul 2023.

82B. Jing, G. Corso, J. Chang, R. Barzilay, and T. Jaakkola, “Torsional diffusion for molecular conformer generation,” arXiv preprint arXiv:2206.01729, 2022.

83I. L. Vazquez-Salazar, E. D. Boißtier, O. T. Unke, and M. Meuwly, “Impact of the characteristics of quantum chemical databases on machine learning prediction of tautomerization energies,” Journal of Chemical Theory and Computation, vol. 17, pp. 4769–4785, July 2021.

84B. Huang and O. A. von Lilienfeld, “Quantum machine learning using atom-in-molecule-based fragments selected on the fly,” Nature Chemistry, vol. 12, pp. 945–951, Sept. 2020.

85C. Zeni, A. Anelli, A. Gilletmo, and K. Rossi, “Exploring the robust extrapolation of high-dimensional machine learning potentials,” Phys. Rev. B, vol. 105, p. 165141, Apr. 2022.

86W. Nash, T. Drummond, and N. Bibilis, “A review of deep learning in the study of materials degradation,” npj Materials Degradation, vol. 2, Nov. 2018.

87B. Fang, Z. Hongfu, and R. Shuhong, “Average life prediction for aero-engine fleet based on performance degradation data,” in 2020 Prognostics and System Health Management Conference, pp. 1–6, 2010.

88M. D. Dangut, I. K. Jemniss, S. King, and Z. Skaf, “A rare failure detection model for aircraft predictive maintenance using a deep hybrid learning approach,” Neural Computing and Applications, Mar. 2022.

89E. J. Gardiner, V. J. Gillet, M. Haranycz, J. H. J. Holliday, N. Malim, Y. Patel, and P. Willett, “Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance,” Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 2, pp. 103–114, July 2009.

90R. A. Miranda-Quintana, D. Bajusz, A. Rácz, and K. Héberger, “Extended similarity indices: the benefits of comparing more than two objects simultaneously. part 1: Theory and characteristics,” Journal of Cheminformatics, vol. 13, Apr. 2021.

91R. A. Miranda-Quintana, A. Rácz, D. Bajusz, and K. Héberger, “Extended similarity indices: the benefits of comparing more than two objects simultaneously. part 2: speed, consistency, diversity selection,” Journal of Cheminformatics, vol. 13, Apr. 2021.

92D. G. Krige, “A statistical approach to some basic mine valuation problems on the Witwatersrand,” Journal of the Southern African Institute of Mining and Metallurgy, vol. 52, pp. 119–139, Dec. 1951. Publisher: Southern African Institute of Mining and Metallurgy.

93V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer New York, 2000.

94M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning,” Physical Review Letters, vol. 108, p. 058301, Jan. 2012.

95K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, “Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space,” The Journal of Physical Chemistry Letters, vol. 6, pp. 2326–2331, June 2015.

96J. Behler, “Atom-centered symmetry functions for constructing high-dimensional neural network potentials,” The Journal of Chemical Physics, vol. 134, p. 074106, Feb. 2011. Publisher: American Institute of Physics.

97A. P. Bartók, R. Kondor, and G. Csányi, “On representing chemical environments,” Physical Review B, vol. 87, May 2013.

98J. Weinreich, N. J. Browning, and O. A. von Lilienfeld, “Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation,” The Journal of Chemical Physics, vol. 154, p. 134113, Apr. 2021.

99L. C. Bluem and J.-L. Reymond, “970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13,” Journal...
nal of the American Chemical Society, vol. 131, pp. 8732–8733, July 2009. Publisher: American Chemical Society.

101 D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” Journal of Chemical Information and Computer Sciences, vol. 28, pp. 31–36, Feb. 1988. Publisher: American Chemical Society.

102 S. Riniker and G. A. Landrum, “Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation,” Journal of Chemical Information and Modeling, vol. 55, pp. 2562–2574, Dec. 2015. Publisher: American Chemical Society.

103 C. Bannwarth, S. Ehlert, and S. Grimme, “GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions.” Journal of Chemical Theory and Computation, vol. 15, pp. 1652–1671, Mar. 2019. Publisher: American Chemical Society.
Improved decision making with similarity based machine learning: Applications in chemistry

Supplementary Information

Dominik Lemm$^{1,2}$, Guido Falk von Rudorff$^3$ and O. Anatole von Lilienfeld$^{3,4,5}$

$^1$ University of Vienna, Faculty of Physics, Kolingasse 14-16, AT-1090 Vienna, Austria
$^2$ University of Vienna, Vienna Doctoral School in Physics, Boltzmanngasse 5, AT-1090 Vienna, Austria
$^3$ Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St. George Campus, Toronto, ON, Canada
$^4$ Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada
$^5$ Machine Learning Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

*Electronic address: anatole.vonlilienfeld@utoronto.ca
(Dated: 30 November 2023)
Local machine learning

Local machine learning methods such as k-nearest neighbours are known since the 1950s and have been shown to be increasingly useful when the training data is not evenly distributed and locality effects dominate the model performance. Formally, such weighted neighbourhood schemes can be expressed as:

\[ y_{\text{pred}} = \sum_{i} W(x_q, x_i) y_i \] (S1)

with weighting function \( W \), query representation \( x_q \) and \( x_i \) and \( y_i \) being the representation and label of prospective training points, respectively. In case of k-nearest neighbour regression, the weighting function is often formalized through a query distance based decay, e.g. \( \frac{1}{d} \), weighting the importance of the local query vicinity higher. Similar to k-nearest neighbours, kernel based machine learning algorithms such as kernel ridge regression can also be seen as a form of weighted neighbourhood learning through the use of specific kernel function, e.g. Gaussian or Laplacian kernels. The local importance weighting is then achieved through optimizing the kernel width \( \sigma \) in \( \exp\left(-\frac{||x_i-x_q||^2}{2\sigma^2}\right) \). Thus, kernel functions without locality weighting do not allow for a direct interpretation as a local learning scheme. Moreover, new predictions are obtained through calculated regression coefficients (see section IV.A) rather than directly averaging the labels \( y \). Besides kernel based methods, also bagged decision tree based models, e.g. random forests\(^1\), can be interpreted as a local learning algorithm. Whereas all prior examples relied on weighting through explicit distance calculations, locality in decision tree’s is achieved through recursively partitioning the feature space to group similar features and labels together. In the case of random forests, an ensemble of multiple decision tree’s are bagged together following the assumption that the prediction of many weak learners will be superior over a single tree. A prediction for an unknown query is made by averaging the prediction of all tree’s, thereby employing a similar weighting effect as described by Eq. S1.

Representation Analysis on QM7/QM9

We applied the SML scheme to learning atomization energies of C\(_7\)O\(_2\)H\(_{10}\) constitutional isomers, QM7, as well as QM9 compounds, respectively. Results in Fig.S3 suggest that SML’s superior data efficiency is consistent throughout datasets and representations, respectively. The performance comparison between multiple representations in Fig.S3C indicate a difference in SML offset depending on the representation. Interestingly, a low offset does not necessarily indicate a higher model accuracy at larger training set sizes. For example, the BoB SML curve in Fig.S3C has a similar offset as FCHL19, however, converges much earlier compared to a better performing representation such as SLATM. Similar learning curve offsets of the global representation BoB\(^2\) and FCHL19 indicate that a combined use of a global representation for fast similarity searches and a local representation for regression might be advisable. In analogy to the learning curve and neighbour analysis performed for FCHL19 in Fig.3B-D, we analysed two compounds predicted with BoB SML models (Fig.S5). Results suggest that nearest neighbours obtained via BoB resemble more chemical similarities than neighbours obtained with FCHL19 (Fig.3D), with BoB even being able to identify a tautomere as the closest neighbour.

FCHL19 Kernel Analysis

Due to the prevalent use of the Gaussian kernel with the FCHL19 representation, we compared the performance of Laplacian and Gaussian kernels. Fig.S6 indicates better performance of a Gaussian type kernel at larger training set sizes. Comparing the kernel principal component analysis (PCA) of a local Laplacian and Gaussian FCHL19 kernel displays a well separated pattern for the Gaussian type kernel (Fig.S7), which could be an explanation for the performance difference. To calculate nearest neighbour distances and similarities with FCHL19, a global Laplacian or Gaussian type kernel is required as defined by Eq.10. The histogram of kernel distance distributions indicates a much wider distribution of similarity values in a Laplacian kernel compared to the Gaussian kernel (Fig.S8), with most of the values being larger than 0.9. The kernel PCA of both kernels, however, shows no clear difference (Fig.S9). Both kernels show a buried PCA pattern for query compounds with high nearest neighbour density (Fig.S9A-B), whereas for a low nearest neighbour density compound the pattern is more separated (Fig.S9C-D). Moreover, comparing the PCA patterns of SML and RML kernels, the former shows more systematic trends, whereas the latter is more randomly distributed (Fig.S10).
**Shared SML Models for Multiple Queries**

To determine the data efficiency of SML in consideration of multiple queries, we train shared neighbourhood SML models by combining the local nearest neighbours of multiple queries to compile the total training set (Fig.S13). Thus, the number of nearest neighbours per query $N_{\text{neigh}}$ becomes a hyperparameter. Moreover, we introduce an additional hyperparameter $N_{\text{filter}}$ that only considers nearest neighbours which occur at least 1, 2, 3 or 4 times in a finite radius of multiple query neighbourhoods. We apply this scheme to train shared SML models tailored for up to 1’000 boutique compound queries of the Enamine REAL library (Fig.S14). Results suggest that for all considered $N_{\text{queries}}$ and $N_{\text{neigh}}$ per query, a combination can be found that reaches 1 kcal/mol chemical accuracy with less than 2’000 training points (Fig.S15). Furthermore, optimal hyperparameter combinations are mostly found for $N_{\text{filter}} = 1$, underlining the importance of neighbour proximity. For example, given two queries, a neighbour within a large radius of both queries could be found ($N_{\text{filter}} = 2$) which might be dissimilar to both queries and, therefore, be less impactful for learning than a direct neighbour ($N_{\text{filter}} = 1$), respectively. Considering the optimal training set sizes for multiple queries (Fig.S15), the time and cost estimate depicted in Fig.5C can be derived.

**SUPPLEMENTARY TABLES**

**TABLE S1.** Overview of maximum training set size $N_{\text{max}}$, number of queries $N_{\text{queries}}$ and representation/similarity metric combination used to identify nearest neighbours for each dataset, respectively.

| Dataset       | $N_{\text{max}}$ | $N_{\text{queries}}$ | Representation | Similarity Metric |
|---------------|------------------|----------------------|----------------|------------------|
| QM7           | 5’680            | 1’421                | Coulomb Matrix | L1-Norm          |
| QM9 C$_7$O$_2$H$_{10}$ | 5’000        | 1’095                | Coulomb Matrix | L1-Norm          |
| QM9           | 133’685          | 200                  | Coulomb Matrix | L1-Norm          |
|               |                  |                      | BoB            | L1-Norm          |
|               |                  |                      | SLATM          | L1-Norm          |
|               |                  |                      | FCHL19         | Gaussian Kernel  |
| Enamine REAL  | 124’440          | 1’000                | Morgan Fingerprint | Tanimoto        |
FIG. S1. Toy example highlighting the effect of edge artifacts on intrinsic dimensionality ($d_{id}$) estimation. (A) Example of a 2D space with evenly spaced points (blue) and fixed boundaries. Given a fixed radius, a point centered in the middle will have more nearest neighbors than a point offset to the boundary. The latter leads to an underestimation of the $d_{id}$. (B) $d_{id}$ of centered and offset points in a fixed boundary 2D space. (C) $d_{id}$ of centered and offset points in a fixed boundary 4D space.
FIG. S2. Randomly chosen atomization energy learning curves of 100 QM9 queries predicted via RML and SML using FCHL19, respectively. The FCHL19 accuracy represents the MAE of an RML model at maximum training set size.
FIG. S3. Learning curves of SML applied to learning atomization energies of C$_7$O$_2$H$_{10}$ constitutional isomers, QM7 and the QM9$^3$ dataset using different molecular representations, respectively. (A) Comparison between SML and RML learning curves applied to C$_7$O$_2$H$_{10}$ isomers using the Coulomb Matrix$^3$ (CM) representation. (B) Comparison between SML and RML learning curves applied to the QM7 dataset using the Coulomb Matrix$^3$ (CM) representation. (C) Comparison between SML and RML learning curves applied to QM9$^3$ using various representations.
FIG. S4. Learning curve of SML applied to learning atomization energies of QM7 compounds using the Coulomb Matrix \(^5\) molecular representations with a random forest regressor.
FIG. S5. Analysis of the influence on neighbourhood density on learning atomization energies of two QM9 compounds with either low or high nearest neighbour density using the BoB representation, respectively. (A-B) Learning curves of SML models for predicting two compounds with either low or high nearest neighbour density, respectively. (C) Amount of nearest neighbours found within an increasing distance radius, respectively. (D-E) Display of the first shell (closest ten molecules) and second shell (next closest fifteen) neighbours for the two query compounds.
FIG. S6. Performance comparison between a Laplacian and Gaussian FCHL19 kernel in combination with SML or RML to learn atomization energies of QM9 compounds, respectively.
FIG. S7. Comparison between kernel PCA’s of local Laplacian and Gaussian SML kernels using the FCHL19 representation applied to a subset of QM9, respectively. (A) SML Laplacian kernel PCA of a compound with high neighbour density. (B) SML Gaussian kernel PCA of a compound with high neighbour density. (C) SML Laplacian kernel PCA of a compound with low neighbour density. (D) SML Gaussian kernel PCA of a compound with low neighbour density.

FIG. S8. Comparison between the distance/similarity distribution of a global Gaussian or Laplacian kernel using the FCHL19 representation on a subset of QM9, respectively. (A) Histogram of kernel similarities of a global Gaussian and Laplacian kernel. (B) Heatmap of sorted kernel similarities per query using a Laplacian kernel. (C) Heatmap of sorted kernel similarities per query using a Gaussian kernel.
FIG. S9. Comparison between kernel PCA's of global Laplacian and Gaussian SML kernels using the FCHL19 representation applied to a subset of QM9, respectively. (A) SML Laplacian kernel PCA of a compound with high neighbour density. (B) SML Gaussian kernel PCA of a compound with high neighbour density. (C) SML Laplacian kernel PCA of a compound with low neighbour density. (D) SML Gaussian kernel PCA of a compound with low neighbour density.
FIG. S10. Comparison between kernel PCA’s of global Laplacian and Gaussian RML kernels using the FCHL19\textsuperscript{4} representation applied to a subset of QM9\textsuperscript{3}, respectively. (A) Local Laplacian kernel PCA. (B) Local Gaussian kernel PCA. (C) Global Laplacian kernel PCA. (D) Global Gaussian kernel PCA.
FIG. S11. Distance based analysis of SML models of learning atomization energies of QM9 using FCHL19 and a Laplacian kernel. $r_{crit}$ represents the distance radius at which a certain percentage of the maximally attainable predictive accuracy is reached. (A) Systematic learning of atomization energies using only training instances within the L1-Distance specified on abscissa. (B) Averaged number of nearest neighbours around a query at different total $N_{max}$ for increasing L1-distance radius. (C) Percentage of data set size necessary to reach a certain accuracy as determined via $r_{crit}$. (D-F) $r_{crit}$, density and intrinsic dimensionality ($d_{id}$) of the QM9 chemical space as a function of available training points, respectively.
FIG. S12. Histogram of calculated angles between 10'000 QM9\(^3\) BoB\(^2\) representation vectors (approximate vector length $\sim$1'000) chosen at random or by similarity, respectively. As a comparison, angles between 10'000 random uniform vectors (vector length 1'000) are shown in red.
FIG. S13. Visualization of the total number of training points $N_{\text{train}}$ as a result of sharing neighbourhoods between queries in the Enamine REAL dataset. To construct a shared SML model for an anticipated number of queries $N_{\text{queries}}$, a certain number of nearest neighbours $N_{\text{neighs}}$ per query have to be chosen to compile a training set. To reduce $N_{\text{train}}$ and $N_{\text{neighs}}$, an additional filter ($N_{\text{filter}}$) can be considered to only include neighbours that occur at least 1, 2, 3 or 4 times in multiple query neighbourhoods.
FIG. S14. Analysis of the optimal combination of $N_{\text{neighs}}$ and $N_{\text{filter}}$ per query in order to reach chemical accuracy (1 kcal/mol) in learning free energies of solvation of queries in the Enamine REAL dataset via SML. $N_{\text{filter}}$ only includes neighbours that occur at least 1, 2, 3 or 4 times in multiple query neighbourhoods. The upper row depicts the boundary for number of training points when an SML would be more efficient than RML. The latter reaches chemical accuracy at $N_{\text{train}}=2'000$ (see Fig.4B). The lower row depicts shared SML model accuracy resulting from a grid search through various $N_{\text{queries}}$, $N_{\text{neighs}}$ and $N_{\text{filter}}$ combinations below the boundary of $N_{\text{train}}=2'000$. 
FIG. S15. Total number of training points needed to generate free energy of solvation SML/RML models that reach chemical accuracy for Enamine REAL compounds as a function of number of queries.
1. Y. Lin and Y. Jeon, “Random forests and adaptive nearest neighbors,” *Journal of the American Statistical Association*, pp. 101–474, 2002.

2. K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, “Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space,” *The Journal of Physical Chemistry Letters*, vol. 6, pp. 2326–2331, June 2015.

3. R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, “Quantum chemistry structures and properties of 134 kilo molecules,” *Scientific Data*, vol. 1, pp. 1–7, Aug. 2014. Number: 1 Publisher: Nature Publishing Group.

4. A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, “FCHL revisited: Faster and more accurate quantum machine learning,” *The Journal of Chemical Physics*, vol. 152, p. 044107, Jan. 2020. Publisher: American Institute of Physics.

5. M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning,” *Physical Review Letters*, vol. 108, p. 058301, Jan. 2012.