Comparison of Active Learning Methods on ANI-1 molecules data set

Luca Viano, 294418, luca.viano@epfl.ch[1]

[1]Master in Computational Science and Engineering, EPFL

"Dubium sapientiae initium" (René Descartes)

***Abstract*— Active Learning (AL) is a promising data science field investigating strategies to find the data whose label is more important to get accurate predictions. In this work, six AL techniques are applied on a data set containing molecules, represented by SOAP [1], aiming to predict formation energies of small molecules. After providing a theoretical background to the implemented models, their performances are tested against a random selection baseline. It arises that the composition of the available pool determines the capability of Active Learning to build an informative subset of the available whole pool.**

I. Introduction

The potential needed to perform a molecular dynamics simulation can be calculated by computationally affordable approximations of the Schrondinger many-body equation like DFT (Density Functional Theorem). Although, considering that the DFT computational cost is still very consistent, there is a concrete interest in finding an alternative model. For this reason, Machine Learning has recently emerged as suitable way to get faster a potential almost as accurate as the one computed by DFT. The speed of the ML computation is strictly related to the amount of data needed in the training set to reach the desired accuracy on the predicted potential. A suitable way to reduce the training set dimension is Active Learning. Active learning roots can be found in the statistics concept known as Optimal Experimental Design already described much before the current "Machine Learning Age" by different authors, e.g. Federov, 1972 [7]. A rebirth of the topic in recent years is due to the need of tools to deal with large data sets whose element labels are scarce and expensive to obtain. This is kind of the situation to face when a regression model for predicting formation energies of molecules is desired. Indeed, Machine Learning methods should rely on DFT as least as possible to become the new paradigm in computational material science and to effectively speed up calculations. This project aims to investigate the efficiency of different Active Learning (AL) technique in building a smaller data set from ANI data set [15] preserving the most informative data.

A. Presentation of the data

The success of all the existing machine learning approach relies on the availability of features as much correlated as possible to the prediction target. Considering that in the situation of this work the targets are the formation energy of the different molecules described in the ANI data set [15], one should find a set of features effectively describing the molecular environment. Smith et al. in [14] exploited a modification of the Behler-Parrinello functions [2] in order to increase the capability of the Machine Learning Model to distinguish various molecular features such as bonding patterns and functional groups. This work is instead based on the SOAP representation proposed by Bartok and Csanyi in [1]. In the context of ML force fields, it consists in building a kernel matrix with structural information as the covariance between atom centred energies. The energy density of an atom is expressed as a superposition of Gaussians projected on a set of basis functions from which rotationally invariant features are extracted. The kernel matrix defining the similarity between atomic structures is computed as sum of the covariance related to each couple of local energy densities. In the linear case adopted here, the operation is equivalent to the covariance matrix of the sums of the local energies in the two different sets of atoms allowing to reduce the computational cost of the procedure. Moreover, the labels considered in the regression model consist on molecules formation energy per atom.

II. Model Based Active Learning Algorithms

The main reference for the techniques implemented in this work is the review by Settles that presents 6 different approaches to Active Learning: Uncertainty Sampling (US), Query by Committee (QBC), Expected Error Reduction, Expected Variance Reduction, Expected Model Change and, finally, some variants of these ones presented as Density-Weighted Methods. As far as the author knows, Uncertainty Sampling has been so far developed only for classification problems and so it was not considered. Query by Committee is one of the most widespread active learning technique and it was used also by Isayev et al. to build up the ANI-1x potential [14]. Expected Error Reduction and expected Variance Reduction are close related because the variance is the only error component that can be reduced acting on the model parameters. Consequently, only a Variance based approach was taken into account. Expected Model Change aims to draw the data that maximises the change in model parameters if added to the training set. For what concerns Density-Weighted methods they are not considered exactly as meant by Settles. Indeed, in III, samples density is used not to modify a model based method but rather as a stand alone way of design an Active Learning scheme. In the following subsections each model is explained deeper.

A. Query by Committee

The main idea behind QBC consists on training K models so that different predictions $\{f_k\}_{k=1}^K$ for the energy corresponding to unlabelled data are obtained. In the QBC picture, each of them is seen as a vote inside the committee. The data x^* whose energy prediction causes the highest disagreement among the votes, is the one that deserves more to be added to $\mathscr{L}$ since evidence of the fact that the actual labelled data cannot provide a reliable prediction for x^* arises from the conflicting votes. A measure of disagreement is required to implement this strategy. As suggested in [3] the disagreement d is measured as variance among the predictions, i.e.

$$d(x) = \frac{1}{K}\sum_{k=0}^{K}(f_k(x) - \bar{f}(x))^2 \tag{1}$$

Notice that $f_k(x)$ denotes the prediction of the k^{th} model in the committee for the data input x while $\bar{f}$ is the average of the K predictions arising from the committee. Finally, a strategy to create the K models is needed. Bootstrap was exploited in this project i.e. K subsets with a size equal to $2/3|\mathscr{L}|$ are randomly extracted to train K linear models that are, then, considered as committee members. In particular $K = 3$ was considered. To conclude, equation 2 shows formally the criterion to choose the data x^* to be added to $\mathscr{L}$.

$$x^* = \arg\max_{x\in\mathscr{U}} d(x) \tag{2}$$

B. Expected Variance Reduction

As stated, a querying technique aiming to minimise the variance of a prediction given by a fixed class of models is equivalent to minimise the error over predictions. Inspired by this theoretical consideration (detailed in Appendix A), the Variance Expected Reduction AL models are build to query the data which is expected to reduce the variance the most when added to $\mathscr{L}$. In formula:

$$x^* = \arg\min_x \frac{1}{|\mathscr{U}\setminus(x,y)|}\sum_{x'\in\mathscr{U}} Var(f(x';\mathscr{L}\cup(x,y))) \tag{3}$$

Since a linear model is used in the current study, observe that:

$$\frac{1}{|\mathscr{U}\setminus(x,y)|}\sum_{x'\in\mathscr{U}} Var(f(x';\mathscr{L}\cup(x,y))) = \tag{4}$$

$$\frac{1}{|\mathscr{U}\setminus(x,y)|}\sum_{x'\in\mathscr{U}} Var(\theta^T\mathbf{x}') \tag{5}$$

Settles [12] suggests to perform the minimisation 3 exploiting the Fisher Information Matrix. A widespread approach in Optimal Experimental Design. Introducing the Fisher Score $\nabla x = \frac{\partial}{\partial\theta} logP_\theta(y|x)$ where $P_\theta(y|x)$ is the probability distribution of the label y given the input x, the Fisher Information F is defined as variance of the Fisher score and so is proportional to the quantity $\sum_x \nabla x \nabla x^T$. The sum index spans all the data available to estimate θ and represents the amount of information they contains about the unknown parameters of the model. Exploiting the Cramér-Rao lower bound, it is known that the inverse of the Fisher Information represents the lowest achievable variance for an estimator $\hat{\theta}$ of the unknown parameter θ. Furthermore, the asymptotic properties of the Maximum Likelihood Estimator provides that the Cramér-Rao lower bound is reached by the MLE variance in the limit of infinite number of elements in the data set. Consequently, when the MLE is considered the minimisation of the Variance can be thought as maximisation of the Fisher information F. This justifies an active learning strategy based on picking the data that, when added to $\mathscr{L}$, maximises F.
When a Gaussian linear model is considered, since $logP_\theta(y|x)$, it follows that:

$$\nabla x = \sum_{\mathbf{x}\in\mathscr{L}} (-y + \theta^T\mathbf{x})\mathbf{x} \tag{6}$$

Observe that since $\mathbf{x}\in\mathbb{R}^D$, each term in the sum is a D-dimensional vector and, so, it is the the Fisher score ∇x. The weights inside the sum are instead the difference between the true label y an the model prediction. Since the proper framework has been established, the Fisher based Active Learning technique can be introduced. Consider the Fisher score of the labelled set plus a point $\mathbf{z}$ taken from the unlabelled data set, i.e.

$$\nabla x^{+\mathbf{z}} = \sum_{\mathbf{x}\in\mathscr{L}} (-y + \theta^T\mathbf{x})\mathbf{x} + (-\hat{y}_z + \theta^T\mathbf{z})\mathbf{z} \tag{7}$$

The true label associated to $\mathbf{z}$ is unknown, so its bootstrap approximation $-\hat{y}_z$ is plugged in to get an approximated fisher Score for the labelled data plus one unlabelled point. We can now query the data that maximises the quantity $F^z = \nabla x^{+\mathbf{z}} \nabla x^{+\mathbf{z}T}$ but because it is a matrix it is still not clear what should be maximised. In [12] three approaches are proposed, D - optimality i.e. minimising the determinant of F^z inverse, A - optimality i.e. minimising the trace of F^z inverse and E - optimality i.e. minimising the highest modulus eigenvalue of F^z inverse. The F^z matrix one gets when the SOAP representation of molecules is considered has eigenvalues very small in modulus so in order to apply D-optimality and A-optimality without overflows the logarithm of the determinant and the trace have been respectively considered. In addition, since the elevated number of features in the SOAP representation makes the Fisher matrix cumbersome to invert, its approximation with its diagonal proposed in [13] was considered.

C. Expected Model Change

The Expected Model Change framework is inspired by the Stochastic Gradient Descent widely used in training ML models. It views the gradient as linear combination of contributes arising from each data in the data set considered individually. Indeed for a generic training set containing N individuals (y_n, x_n), the Mean Square Error (MSE) gradient can be written as sum of gradients of the single elements

square error in the following denoted as ∇MSE_n

$$\nabla MSE = \nabla \frac{1}{2N} \sum_{n=1}^{N} (y_n - \theta^T x_n)^2 \tag{8}$$

$$= \frac{1}{2N} \sum_{n=1}^{N} \nabla (y_n - \theta^T x_n)^2 \tag{9}$$

$$= \frac{1}{2N} \sum_{n=1}^{N} \nabla MSE_n \tag{10}$$

The fundamental observation leading to Expected Model Change is that data x_n with a large associated MSE_n would contribute significantly to the gradient if considered for training the model, i.e. they would cause a consistent change in the parameters. The argument provides evidence in favour of an AL strategy based on querying the point whose loss function gradient norm $||\nabla MSE_n||$is the largest. The method has been implemented following [4], i.e. we pick x^* so that:

$$x^* = \arg\max_{x \in \mathscr{U}} \frac{1}{K} \sum_{k=1}^{K} ||(f(x) - y_k)x|| \tag{11}$$

It is important to observe that $\{y_k\}_{k=1}^K$ is a set of prediction for the unknown label of x obtained by a bootstrap technique based on K different models. In the following this scheme is referred as ECNA (Expected Change Norm Average) and it is compared to a computationally more attractive variant that it is defined here as ECLA (Expected Change Labels Average). ECLA criterion selection reads as follows:

$$x^* = \arg\max_{x \in \mathscr{U}} \left(f(x) - \frac{1}{K} \sum_{k=1}^{K} y_k \right) x \tag{12}$$

III. Model Free Active Learning Algorithms

According to [9], an algorithm is defined model free if it does not require the labels of $\mathscr{L}$ to assess which sample in $\mathscr{U}$ deserves more to be added to the data set. An algorithm of this family usually exploits a measure coming from a balance between similitude and diversity between the data. Two methods of this type were considered: EGAL based on both similitude and diversity and FPS purely based on diversity. Considering that depicting the data point in the design matrix columns space, in the following "distant enough" is used as a synonym of "diverse enough" and the similarity between a data and the others is associated to the density around that point.

A. FPS: Farthest Point Sampling

FPS approach can be seen as purely diversity based algorithm. During the active learning procedure, given the current labelled set $\mathscr{L}$, it samples the points from the rest of the data set whose minimum distance from the point in $\mathscr{U}$ is the highest. In order to quantify the distance, it is necessary to introduce a kernel function $k(\cdot,\cdot) : \mathbb{R}^{dim(x)} \times \mathbb{R}^{dim(x)} \to \mathbb{R}$ and the induced normalised kernel

$$K(x, x_L) = \frac{k(x, x_L)}{\sqrt{k(x,x)k(x_L,x_L)}} \tag{13}$$

At this point the distance measure can be defined as $D(x, x_L) = \sqrt{2 - 2K(x, x_L)}$ and formalise the choice criterion in formula as follows:

$$x^* = \arg\max_{x \in \mathscr{U}} \min_{x_L \in \mathscr{L}} D(x, x_L) \tag{14}$$

What it has been described so far holds for each suitable choice of the kernel function. In this work it was picked simply equal to the scalar product so that the kernel corresponds to the cosine similarity matrix. Finally, the AL method proposed here is the same used in [5] in the context of dimensionality reduction. In this case the selection criterion is the same but the distance is the one between two features seen as points in the row space of the design matrix X. In IV-C the FPS method is applied following the latter point of view.

B. EGAL : Exploration Guided Active Learning

EGAL is a generalisation of FPS that considers also the density of points in the column space in the Active Learning selection. This idea is implemented setting a similitude threshold under which a data is deemed different enough from the already labelled one. Then, among the data selected in this way the one lying in the most dense region is added. This attempts to avoid the selection of leverage points i.e. the ones that are very different from the others and isolated in the column space. They could, indeed, lead to an unbalanced design matrix and, consequently to a worst fit. More in details, the similarity matrix of all the data is computed at first. To this purpose the scikit learn [10] function *cosine_similarity* was used. Afterwards, one defines the neighbours set $\mathscr{N}_i$ as set of point similar enough to the data x_i i.e. points such that the cosine similarity $sim(\cdot,\cdot)$ is above a threshold α. Density is so measured as $density(x_i) = \sum_{x_r \in \mathscr{N}_i} sim(x_r, x_i)$. Notice that the density matrix defined above does not depend on which data are currently labelled or unlabelled. Thus, it is computed only once at the beginning and then stored in memory. It is not the same for the diversity measure. Since it aims to understand how different an unlabelled point x_i is from the ones in the labelled data set it is quantified as inverse of the highest similitude score between x_i and the labelled set. At each step a candidate set CS is so defined using a threshold β to require enough diversity.

$$CS = \left\{ x_i \in \mathscr{U} \,|\, \arg\max_{x_r \in \mathscr{L}} sim(x_i, x_r) \leq \beta \right\}$$

Then if one chooses to add only one point per step to $\mathscr{L}$, the one whose label is required according to EGAL scheme is:

$$x^* = \arg\max_{x \in CS} \quad density(x) \tag{15}$$

If, instead, K points should be added, the K highest density scores are considered. Thresholds choice is fundamental to decide whether to balance the algorithm in favour of diversity or density. A low β shrinks the candidate set privileging the diversity. Its choice was done following the procedure outlined in EGAL's authors paper [8].

IV. Computational considerations for the solution of the regression problem

A drawback common to the a majority of Active Learning ideas is the computational complexity. Indeed, the Model Based techniques require at least the solution of a linear system to get the ridge weights (Equation 16).

$$(X^T X + \lambda I)\theta = X^T y \tag{16}$$

In addition, more of them have to be solved when a bootstrap ensemble is considered. In order to face the problem 2 different approaches have been taken into account: the use of an iterative solver and the reduction of the X features number.

A. Iterative Solver

One can notice that the systems solved at two consecutive steps differs only because of the addition of a certain number of new samples N_{new} to the design matrix X, i.e. the one sampled by the active learning routine at the previous step. The optimal weights θ are consequently expected to be close between two consecutive steps. Hence an iterative method as gradient descent with momentum (see IV-B) appears attractive because if the algorithm converged close to the optimal weights at one step, this point in the error function landscape can be used as initial position at the next call of the gradient descent routine. With this choice the convergence at the next step takes by far less iterations. The discriminant to determine the feasibility of the iterative solver is the number of iteration needed to convergence at the first iteration when no information about a preferable starting point is available. As target to assess the convergence, the solution of the system obtained by LU decomposition followed by backward substitution (i.e. the output of Numpy *solve* method) was considered.

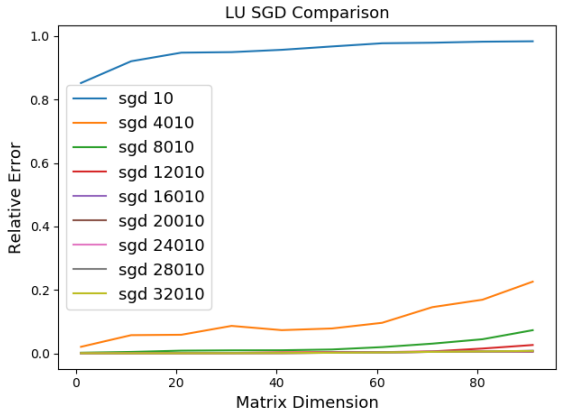

Fig. 1: Norm of the difference between SGD and LU solutions against dimension of the matrix. The iterative method used was the momentum modified SGD with parameters batch size = 20, $\alpha = 0.9989$ and $\gamma = 0.01$. On the y-axis the quantity is dimensionless since is the ratio between the norm of the difference and the norm of the LU solution

The Euclidean norm of the difference vector between the sgd and numpy result is plotted against the number of new samples added to the matrix in Figure 1. The method well behaves in the limit of a large number of iterations since a curve closer and closer to zero is observed but the number of iteration needed already for a small initial pool size between 10 and 100 samples does not suggest to apply it during the active learning procedure. Basically, it is true that it becomes less and less expensive once the algorithm get closer to the optimal solution but the cost of reaching that area is too high and spikes in favour of the LU based approach. It can be accelerated avoiding to build the kernel from scratch at each iteration. Indeed, the matrix $X^T_{+z}X_{+z}$ updated with the sample $z = [z_1, \ldots, z_D]$ is conveniently computed observing that $X^T_{+z}X_{+z} = X^T X + z^T z$.

B. SGD with momentum

As a side note from the active learning scenario, the main focus of this paper, a brief explanation of the considered iterative method is here provided. Let θ^k be the weights at the iteration k of the algorithm, γ be the learning rate and $\mathscr{L}$ be the objective function. In standard SGD the update rule reads:

$$\theta^k = \theta^{k-1} - \gamma \nabla_\theta \mathscr{L}$$

The convergence in this case results to be slow, momentum method is a common way to accelerate the convergence. It exploits the step done between θ^{k-1} and θ^k to compute the update step from θ^k to θ^{k+1}. Denoting the update step $\Delta^k = \theta^k - \theta^{k-1}$, the update rule takes the form:

$$\Delta^k = -\gamma \nabla_\theta \mathscr{L} \quad k = 1$$

$$\Delta^k = -\gamma \nabla_\theta \mathscr{L} + \Delta^{k-1}\alpha \quad k \neq 1$$

The observed effect in literature [11] of this method are a shorter path towards the minimum and an induced increase in the learning step when some consecutive steps are almost parallel. They both are beneficial.

C. Feature Selection

A feature selection process is necessary to speed up the solution of the linear systems involved in the various method. To pursue this aim three techniques have been investigated. These are the F-score selection scheme, Farthest Point Sampling used this time on feature not on samples as before and PCA. Ten thousand conformers from different molecules were considered to build a full-feature design matrix X based on SOAP. Observe that $X \in \mathbb{R}^{(10000 \times 4032)}$.The linear regressor trained on X reaches the benchmark error on the COMP6 data set showed with the black solid line in Figure 2. In order to assess the best scheme we select the best K features from X obtaining a matrix $X_{reduced} \in \mathbb{R}^{(10000 \times K)}$ for each one of the three methods. The error on the same test set to get the RMSE of the full model was used to evaluate the error of the reduced models. In the loglog error plot shown in Figure 2 shows how for very small number of features PCA is the best method since it can pick linear combinations of the original features. Despite being forced to select only original features FPS reaches PCA performances when more than 100 of features are kept in the model and it is by far best of the F

regression selection. Furthermore, it can be noticed that with 500 features in the model one gets an order of magnitude in the RMSE equal to the one attained by the full model. As a consequence, the simulation of the different active learning techniques relies on the reduced matrix with 500 features given by the FPS algorithm.

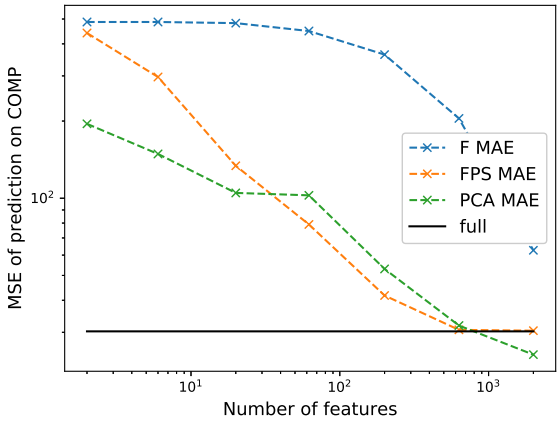

Fig. 2: Comparison of feature selection techniques ANOVA and FPS with PCA (feature reduction technique)

V. Results

The computational experiment consists in building a labelled subset, denoted as $\mathscr{L}$, including $|\mathscr{L}|$ samples retaining as much information as possible from an initial set of N unlabelled molecules. In practice, it means to seek $\mathscr{L}$ that allows to a linear model trained on it to minimise the RMSE of the energies predicted for the molecules in a given test set. Two different scales are considered: all the methods have been tested with $N = 20000$ and $|\mathscr{L}| = 5000$, then the fastest ones are evaluated also at a larger scale $N = 400000$ and $|\mathscr{L}| = 100000$.

A. Choices of the initial unlabelled subsets of ANI

It is expected that the diversity among the data of the initial unlabelled data set influences the results of the AL techniques comparison. A data set characterised by a low diversity is highly redundant i.e. it includes many molecules that are conformers. In order to verify this assumption on the smallest scale, 3 different unlabelled pools with a size $N = 20000$ molecules have been prepared. They are outlined below following an order of increasing redundancy:

- **Pool A**. It is designed to avoid redundancy in the sense that it contains only different molecules. In ANI there are 57462 groups of conformers. Consequently, Pool A is created picking one molecule for each one of this groups obtaining in this way a set of 57462 molecules (Pool D). Finally, the desired size $N = 20000$ molecules is obtained randomly picking N molecules from pool D.
- **Pool B** It is obtained by considering the union of groups of conformers of the same molecule. Consequently, once the groups size is fixed to S, N/S group indices are extracted from Pool D. Finally S molecules are selected from each one of the indices. The union of all these subsets gives the desired size $N = 20000$. In the specific case of this work S is set to 80. It follows that redundancy is present but it is unbiased i.e. all the conformers groups are represented by the same amount of samples.
- **Pool C** It is created focusing on intentionally introducing redundancies biased towards not informative molecule. It is constructed joining 2 very different sets that can be respectively associate to an highly and scarcely degree of diversity. On the one hand, the first half was selected following the approach used in designing pool A. Indeed, $N/2 = 10000$ molecules are randomly picked from ANI. On the other hand, the second half contains only conformers taken from the ANI molecules with 1 or 2 heavy atoms. It is evident that the latter half does not contain informative samples since COMP6 contains molecules with at least 7 heavy atoms so it is expected that this redundancy are avoided by a suitable Active Learning technique. The redundancy in pool C is thus biased towards small molecules.

The different composition of the described pools can be appreciated thanks a 2D PCA projection, noticing that the conformers lie close each other in the plot. Consequently, clusters are observed in the PCA of Pool B and Pool C, i.e. in the pools where conformers have been voluntarily introduced.

B. Description of the test sets

As introduced, the quality of a labelled subset is assessed by computing the RMSE and MSE on 4 different sets:

- **U** It is just the unlabelled pool, i.e. the points that are still unlabelled at some point along the selection process. Observe that it changes along the Active Learning execution.
- **ANI3000** A set of 3000 different molecules taken from ANI
- **COMP1500** A set containing a molecule for each of the 1500 groups of conformers in the COMP6 [14] with 7, 8 or 9 heavy atoms.
- **COMP3000** A set containing a molecule for each of the 3000 groups of conformers in COMP6. It is important to notice that COMP3000 includes molecules with up to 13 heavy atoms while in ANI the largest molecules have 8 of them.

The idea behind taking into account 4 test sets is to assess the performance in case of molecules more and more different from the ones in the initial unlabelled pool. Indeed, it is observed in [9] that the Active Learning methods usually manage to create a better subsets than the random sampling strategy when the unlabelled pool at each iteration is considered as testing set but this does not guarantee they are the best also when a more general test set is taken into account. As a consequence, the test on the COMP1500

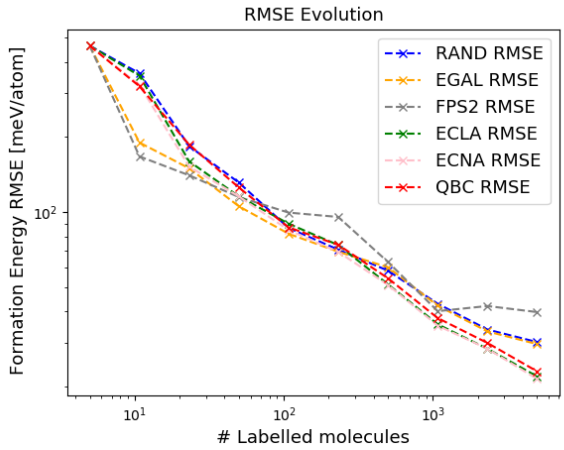

Fig. 3: Root Mean Square Error of the predicted energy on U as function of the selected points from Pool C by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

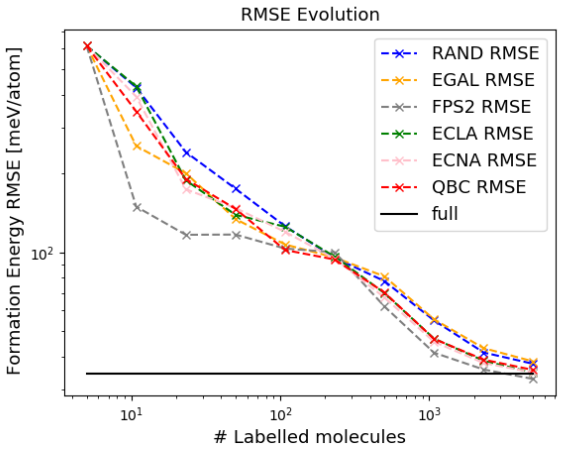

Fig. 5: Root Mean Square Error of the predicted energy on the COMP1500 as function of the selected points from Pool C by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

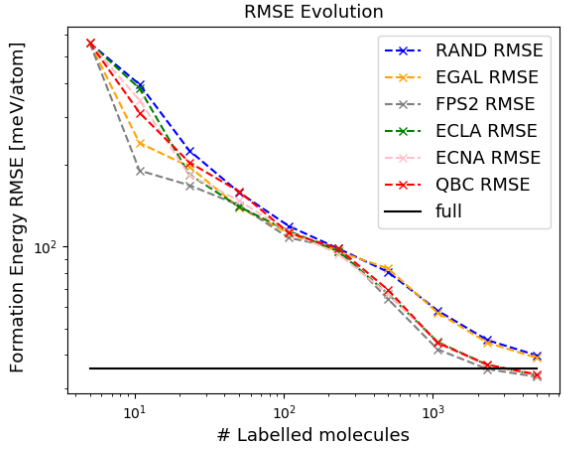

Fig. 4: Root Mean Square Error of the predicted energy on ANI1500 as function of the selected points from Pool C by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

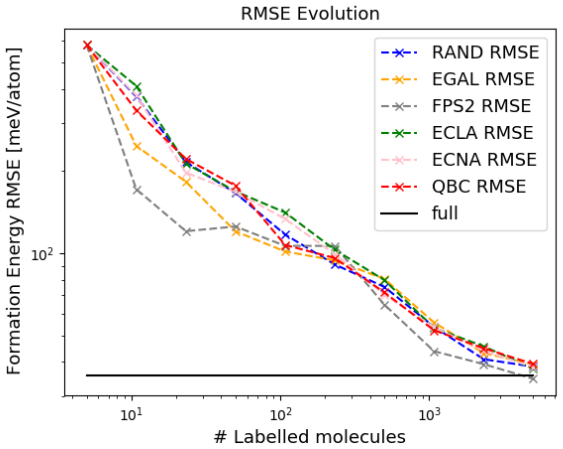

Fig. 6: Root Mean Square Error of the predicted energy on COMP3000 as function of the selected points from Pool C by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

and COMP3000 aims to investigate this. In particular, the performed simulations aim to probe if the active learning selected pools are the best also in the prediction of larger molecules energies.

C. RMSE evolution of the AL techniques on Pool C

The results shown in Figures 3, 4 5 and 6 shows the results of EGAL, FPS, ECLA, ECNA, QBC against the random naive strategy (RAND) when Pool C is considered as initial set.

Using U as a test set, it is observed that the methods exploiting bootstrap (ECLA, ECNA and QBC) are the ones leading to the lowest RMSE. FPS samples very informative molecules till to about 50 molecules but then it is outperformed by the other techniques reaching the worst RMSE at the final considered size. Finally, EGAL performs almost as efficiently as FPS for the smallest considered sizes and slightly better than RAND at the final size. The RMSE chart arising from ANI3000 highlights that all the Active Learning techniques improves RAND RMSE. As in U, bootstrap based methods performs particularly well for large sizes while EGAL is close to FPS for small sizes and near to RAND for large ones. Main difference is that FPS that was the worst over U for large sizes is the best for the energy prediction on ANI. As long as COMP1500 is considered as test set, it is seen that the active learning techniques are always preferable. The one that surprises the most is FPS that manages to pick 20 points data set that have the same performance of the random selected pool with 200 points. When the labelled set assumes a size larger than 1000 molecules, it is again observed that FPS performs the best, that the ECNA, ECLA and QBC reach approximately the same error and, finally, that EGAL attains the same result as RAND. When

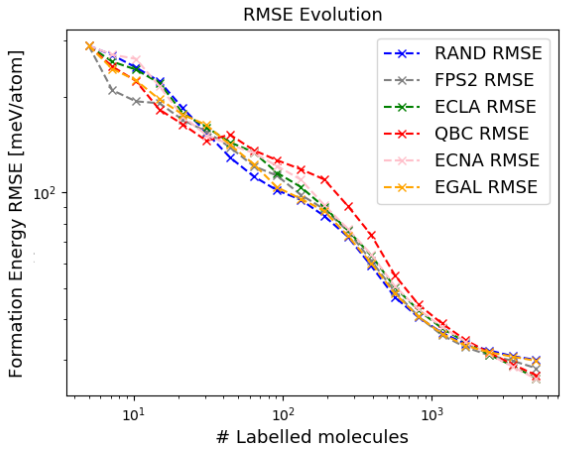

Fig. 7: Root Mean Square Error of the predicted energy on U as function of the selected points from Pool B by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

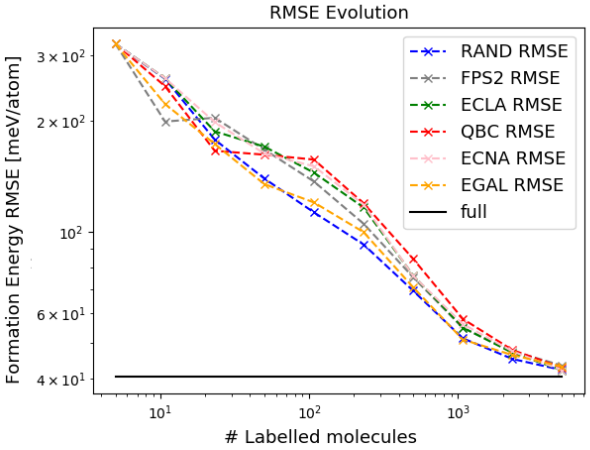

Fig. 9: Root Mean Square Error of the predicted energy on the COMP1500 as function of the selected points from Pool B by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

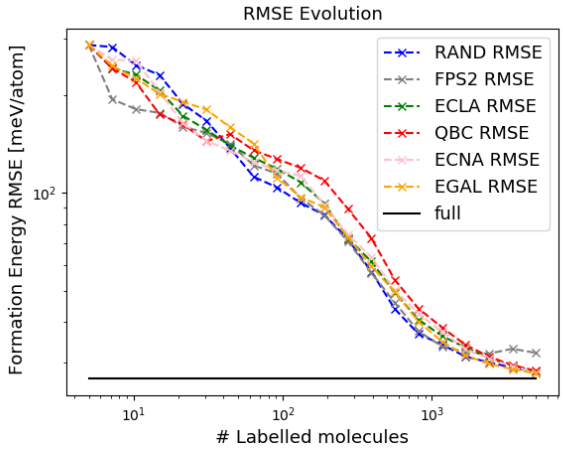

Fig. 8: Root Mean Square Error of the predicted energy on ANI3000 as function of the selected points from Pool B by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

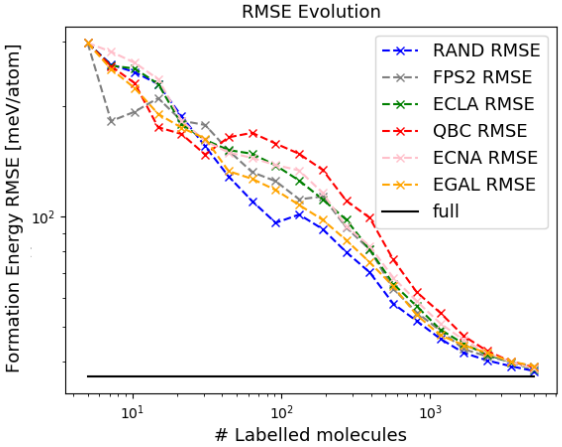

Fig. 10: Root Mean Square Error of the predicted energy on the COMP3000 as function of the selected points from Pool B by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

COMP3000 is considered RAND is no longer the worst. In particular the 3 bootstrap based method are often worst than RAND. It seems that the model free techniques manage to generalise better, indeed FPS and EGAL exhibit the same behaviour shown on test set COMP1500. Comparing the results emerging from Pool C, it is noticed that the methods based on bootstrap perform similarly in all the test set. This suggests that the specific AL criterion applied on the labels computed by bootstrap is not implying a significant change in the efficiency of the method. In addition, considering that FPS remarkably emerges as best techniques in the most diverse set from Pool C (COMP1500, COMP3000), one can suggest that diversity among data is the concept leading to the best Active Learning model in improving the prediction on a more general data set. On the contrary, diversity looks to be detrimental in the case of U. It emerges that, when the test set is similar to the selected data, it can be beneficial to include a density-diversity balance in the selection criterion as implemented in EGAL. Although, Query by Committee and Expected Model Change techniques could be even more favourable in this condition.

D. RMSE evolution of the AL techniques on Pool B

The results achieved by the simulations run on Pool B are reported in Figures 7, 8, 9 and 10. Observing the graph illustrating the models performances on U it is observed that all the AL techniques improves random baseline RMSE for sizes up to 30 molecules and beyond 2000. The only scheme showing a curve significantly different from the others is Query by Committee (QBC). Indeed, it reaches the lowest error in the 2 regimes where RAND is outperformed by AL approaches but for sizes in between the range considered in the simulation it gets the highest RMSE. Once again, it

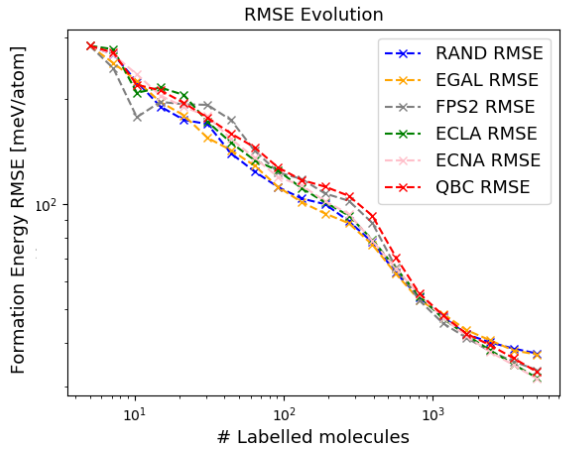

Fig. 11: Root Mean Square Error of the predicted energy on U as function of the selected points from Pool A by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

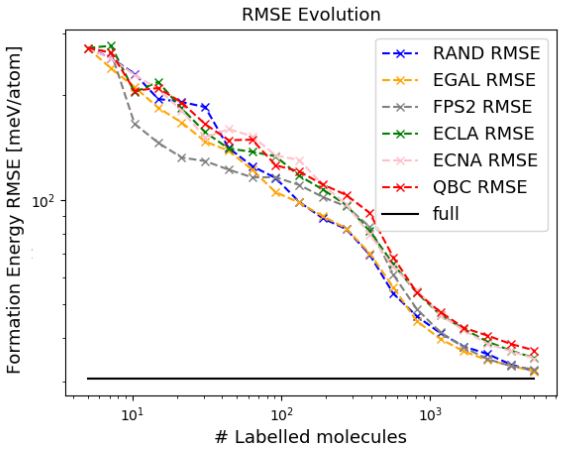

Fig. 13: Root Mean Square Error of the predicted energy on the COMP1500 as function of the selected points from Pool A by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

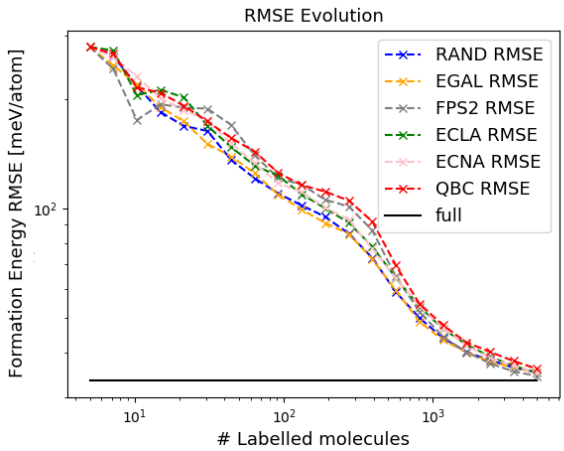

Fig. 12: Root Mean Square Error of the predicted energy on ANI3000 as function of the selected points from Pool A by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

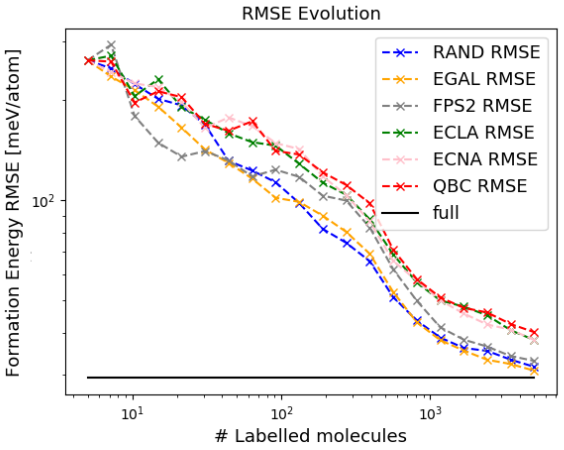

Fig. 14: Root Mean Square Error of the predicted energy on the COMP3000 as function of the selected points from Pool A by the various AL techniques. They come as result of the seed over the choices of 10 possible different starting sets.

is observed that bootstrap based methods reach the lowest RMSE at the largest size considered in the simulation. The results on ANI and COMP are instead less favourable to active learning. Indeed it is seen that apart from small sizes the Active Learning curves lie above the random one in the RMSE plot. Although, it is noticed that the RMSE of the implemented methods decreases faster than the RAND one for sizes larger than 1000 molecules. Consequently, it can be proposed that the initial size of the labelled pool determines the AL schemes efficacy. The fact could be verified running a simulation starting from a labelled set of approximately 1000 molecules. Same conclusions can be drawn considering COMP1500 and COMP3000 as test set that provides even clearer evidence on the impossibility of the AL algorithms to decrease the RMSE of RAND for a size in the range between 10 and 100 labelled molecules.

E. RMSE evolution of the AL techniques on Pool A

Analogously, Figures 11, 12 and 14 shows the RMSEs on the same test sets as before when Pool A is used as available sampling pool. It can be show that in this context the efficiency of the various methods is significantly different from the trend observed for Pool C but aligned with the results on Pool B. Indeed, the bootstrap based techniques reaches the best RMSE on U for a labelled size of 5000 molecules but all the schemes leads to an error higher than the RAND one for sizes in the central part of the considered range. On ANI the only methods that appear to perform in a moderate favourable way are EGAL that attains errors always very close to RAND and FPS that improves the prediction accuracy for the largest considered size. On the contrary QBC, ECLA and ECNA fails in sampling informative points. Interesting results come from the simulation using either

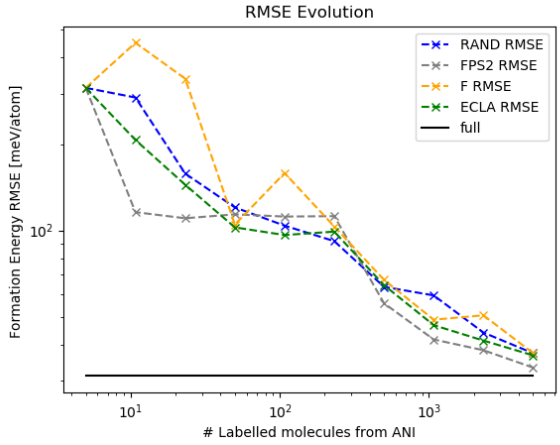

Fig. 15: Root Mean Square Error of the predicted energy on COMP3000 as function of the points selected from Pool C by ECLA, FPS and F. Only 1 seed whose considered due to the slowness of F.

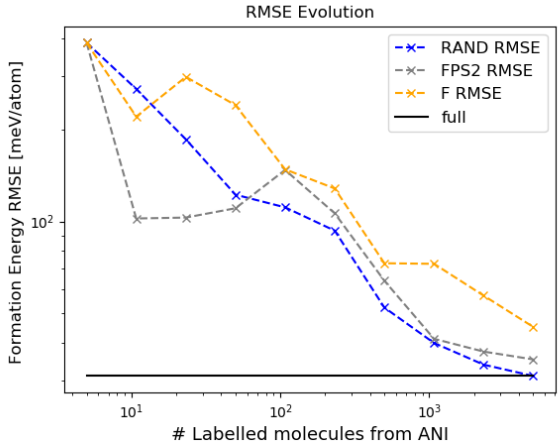

Fig. 16: Root Mean Square Error of the predicted energy on COMP6 as function of the points selected from Pool A by FPS and F. Only 1 seed whose considered due to the slowness of F.

COMP1500 or COMP3000 as a test set. Indeed the bootstrap based techniques perform significantly worst than the others for sizes above 1000 while both the Model Free models surpass RAND selection RMSE for small sizes. Furthermore, EGAL is helpful also for sizes larger than 1000 molecules. FPS shows the behaviour already exhibits in the previous contexts. It improves the RMSE dramatically at the first steps of the AL procedure but then it is surpassed by EGAL and RAND for sizes about 100 molecules.

F. Expected Variance Reduction Results

It remarkably emerges that Fisher Information method (F) severely suffers from instability. Indeed, in Figure 15 and 16 it is observed that the Fisher Information based method behaviour is not monotonously decreasing as the size of the labelled data increases. A plausible way to fix it would be a density weighted correction. In addition, carefulness is needed because the results are based on only 1 seed whereas the high running time of the Fisher method. Despite this fact, it can be concluded that the Fisher based method is probably the one with the most solid theoretical background but when the available data set has a huge dimension, as in this case, it requires the severe approximation of the Fisher Information Matrix with its diagonal. This does not speed up the running time enough to make it attractive in a concrete application and it is probably the approximation that brings the observed oscillation in the learning curves. These issues make Fisher matrix based method unfeasible for this data set.

VI. General Remarks

A. Density correction

The poor performance of many Active Learning techniques on pool A can be probably explained by the fact that since there are not conformers in the available pools all the possible choices are informative enough to improve the regression accuracy. Consequently, the effect of the AL strategy on this data set is no longer taking more informative points but taking points from regions where the data points are less dense. These thus become leverage points i.e. they affect the global behaviour of the linear regression model more than what the other data does in average. Indeed, data that have an extreme value on one or more explanatory features have the effect of forcing the fitted model to pass very close to the observed label of this point [6]. Moreover the PCA representation of the selected points reinforce this interpretation. It is noticed the QBC preference for points lying at the borders of the data pattern. These points influences excessively the linear fitting because no other observations are available in correspondence of that area of the feature space. It follows that the AL predictions are less accurate. Moreover, the poor performance on pool B indicates that on a mildly redundant pool, random selection approach is not improved by Active Learning techniques. The argument based on leverage points holds also in this case. Indeed, in a limit of a labelled pool large enough the random sampling approach leads to a set where the structure of groups counting on the same amount of conformers is still satisfied. This implies a well balanced designed matrix. On the contrary, the data set structure may be distorted by an active learning technique and consequently the molecules in the least represented conformers group act as leverage points. A possible approach to avoid the leverage point sampling is adding a regularisation term to the selection criterion. The lower the density around the data point the more penalising is the introduced term. This approach is proposed in [12] as Density-Weighted methods. A strong hint towards such a remedy arises from the performances generally observed for EGAL. Indeed, it is observed that it never leads to a RMSE significantly worst than the RAND and it is the only method able to outperform RAND in pool A i.e. on the pool that does not contains redundancies. In addition, when it is applied to Pool C it is not as efficient as a purely diversity based method as FPS but, for some sizes, it still manages

to provide labelled sets better than what the bootstrap based methods do. Such a characteristic stability follows from the fact that this is the only technique taking in account a density measure in the decision criterion. It makes the EGAL method more versatile and it seems to guarantee the construction of a good labelled subset whatever is the redundancy degree of the available total pool. This spikes in favour of correcting all the other methods using density to make them more robust. Furthermore, it is necessary to remark that leverage is an issue specific of linear models, consequently an Active Learning techniques that leads to disappointing results when a linear model is used as regression model, can work when a more flexible model (e.g. Kernel regression, neural networks) is considered.

B. Consideration about the running times

As previously described the model based methods requires the Bootstrap estimation of the labels in $\mathscr{U}$ at each step. On the contrary, the model free method needs only to compute a similarity and/or diversity scores to decide which point to add. Even more favourable in EGAL is that the similarity scores do not depend on the composition of $\mathscr{U}$ and $\mathscr{L}$ at each time step so it can be computed once for all before starting the Active Learning procedure. This holds also for the distances in FPS but it does not for the diversities scores in EGAL that must be updated and sorted after each modification of $\mathscr{U}$ and $\mathscr{L}$. In particular, the sorting turned out to be the operation determining the running time of the algorithm. Indeed, in order to run EGAL in a reasonable time the sorting operation has been parallelized. According to the above statements, FPS turned out to be the fastest method. Finally, the expected model change variant defined as ECLA is noticed that to performs consistently with the original proposed ECNA but with a lower running time.

VII. Simulation on larger available pool

So far the implemented techniques have been tested on pools denoted as A, B, C having the same size fixed to 20000 molecules. The fastest method has been tested also on 2 larger data sets:

- **Pool D** It has been already defined in constructing Pool A.
- **Pool E** It is much larger since it contains Pool D as a subset and other conformers randomly picked from ANI molecules with 1,2,3 and 4 heavy atoms. It counts 400000 elements and it embodies the biased redundant data set.
- **Pool F** It also contains Pool D as a subset but it takes K specimens for each conformers group that was represented by only one molecule in Pool D. In the specific case, Pool F has been built considering $K = 80$. It is analogous to Pool B since each conformer groups is equally represented.

Since this work focuses on investigate the AL performance on a test set more general than the available pool, only the error on the test set COMP3000 is outlined in the following. The RMSE plots obtained when Pool D, Pool E and Pool F

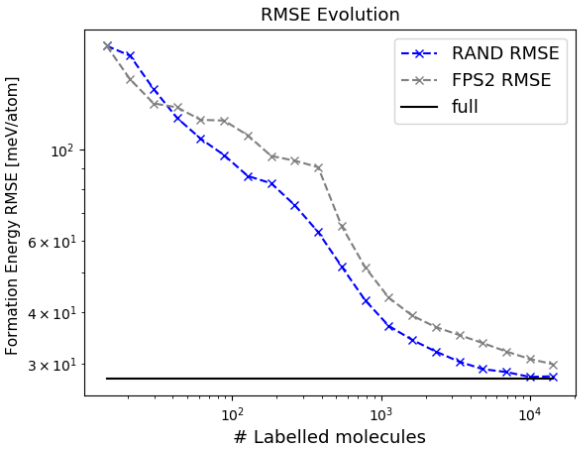

Fig. 17: Root Mean Square Error of the predicted energy on the test set COMP3000 as function of the points selected from Pool D by FPS.

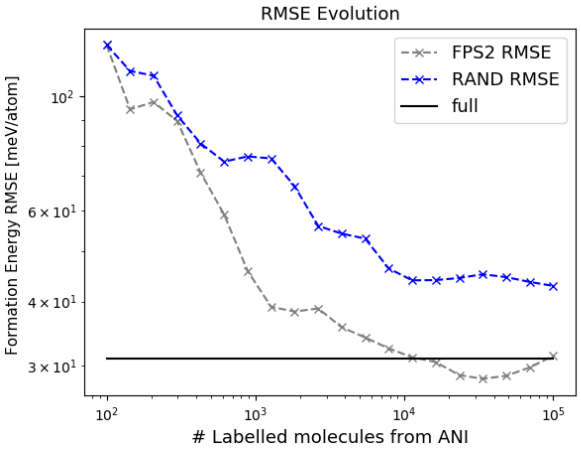

Fig. 18: Root Mean Square Error of the predicted energy on the test set COMP3000 as function of the points selected from Pool E by FPS.

are respectively considered as available pool are reported in Figure 17, 18 and 19. As final size of the available data set it is kept the same criterion used in the simulations concerning available pools containing 20000 molecules. i.e. the final labelled set size is set to a quarter of the size of the available pool. The results are aligned on the ones observed in the smaller pools. In particular one notices that the random baseline performs better when Pool D or Pool F are considered while it is significantly improved when Pool E is used. In particular, it emerges that FPS is not helpful at small sizes when Pool F is considered. On the contrary, comparing the most redundant cases at the 2 different scales (Pool C and Pool E), it is observed that considering a larger pool FPS leads to a larger advantage. Notice, for instance, that FPS manages to select a labelled set containing 10^4 molecules whose RMSE is the same that the full data set manage to reach.

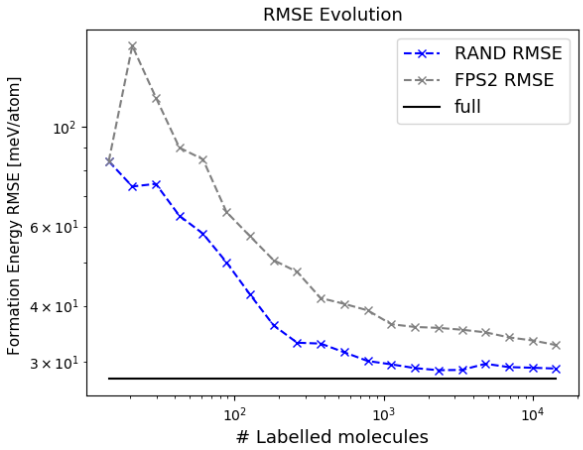

Fig. 19: Root Mean Square Error of the predicted energy on the test set COMP3000 as function of the points selected from Pool F by FPS.

VIII. Conclusion and Future Perspective

It is observed that all the investigated techniques are useful if the available pools presents biased redundancies . If this is not the case, AL techniques in particular the bootstrap based methods turn out to be prone in sampling leverage points that are detrimental in a linear regression scenario. The affirmation of EGAL as most robust technique to this specific drawback suggests to investigate in future the effect of applying a density correction to QBC, ECLA and ECNA. In addition, EGAL has been applied fixing the involved parameters as suggested by the authors. Balancing the diversity density EGAL contributions more in favour for diversity seems to be a promising field to get an efficient active learning techniques whatever is the redundancy degree in the available pool. Moreover, the results obtained here in the specific context of the ANI data set agrees with the more general study carried out in [9] on the better performances reached by the Model Free schemes. In addition, noticing that bootstrap based techniques are often the best on the test set U but the least attractive when COMP is considered, it can be claimed that in the case studied here the RMSE on the unlabelled test used as evaluation criterion in [9] does not imply a good generalisation capability of the model. Finally, the emerging fact that QBC does not really boost a linear model predictive power is reconciled with its efficiency when a neural network is adopted as a model (as in ANI-1X) observing that neural network performance is not affected by an unbalanced design matrix (i.e. leverage points). Consequently, it could be investigated if the labelled set selected by techniques prone to sample leverage points such as QBC could lead to a low RMSE when it is used as training set for a more flexible model. If this holds, methods based on a linear models ensambles could be used as cheap way to select a suitable training set for a computationally more demanding model.

IX. Acknowledgments

My sincere gratitude goes to Félix Musil for its weekly supervision and constant availability.

Appendix

A. Equivalence between maximum error reduction and maximum variance reduction

Assume that the true label y is modelled as $y = f(x) + \varepsilon$ where $f(x)$ is an unknown deterministic function and ε is zero mean random noise. In addition, consider now a class of models $\hat{f}$ that are trained to approximate y. The generalised error of the prediction relative to any data $x \in \mathscr{U}$ is:

$$\mathbb{E}_{\mathscr{L}}\left[(y-\hat{f}(x;\mathscr{L}))^2\right] \tag{17}$$
$$= \mathbb{E}_{\mathscr{L}}\left[(y-f(x)+f(x)-\hat{f}(x;\mathscr{L})^2\right] \tag{18}$$
$$= \mathbb{E}_{\mathscr{L}}\left[(y-f(x))^2\right] + \mathbb{E}_{\mathscr{L}}\left[\left(f(x)-\hat{f}(x;\mathscr{L})\right)^2\right] \tag{19}$$
$$= \mathbb{E}_{\mathscr{L}}\left[(y-f(x))^2\right] \tag{20}$$
$$+ \mathbb{E}_{\mathscr{L}}\left[\left(f(x)-\mathbb{E}_{\mathscr{L}}[\hat{f}(x;\mathscr{L})]+\mathbb{E}_{\mathscr{L}}[\hat{f}(x;\mathscr{L})]-\hat{f}(x;\mathscr{L})\right)^2\right] \tag{21}$$
$$= \mathbb{E}_{\mathscr{L}}\left[(y-f(x))^2\right] \tag{22}$$
$$+ (f(x)-\mathbb{E}_{\mathscr{L}}[\hat{f}(x;\mathscr{L})])^2 \tag{23}$$
$$+ \mathbb{E}_{\mathscr{L}}\left[(\hat{f}(x;\mathscr{L})-\mathbb{E}_{\mathscr{L}}[\hat{f}(x;\mathscr{L})])^2\right] \tag{24}$$

From (17) to (24), the notation $\mathbb{E}_{\mathscr{L}}[\cdot]$ means expectation over all possible training set $\mathscr{L}$ and the dependence of the model $\hat{f}$ on $\mathscr{L}$ is due to the fact that the optimal model parameters depend on the data exploited to train it. Furthermore, observe that from (18) to (19) the zero mean of the noise was exploited. The aim of the previous calculation is showing that the generalisation error can be written as sum of three terms: (22) represents an error that also the true unknown model commits because of the stochastic effects that influence the value of y. In a physical context, it can be imagined that the data describing a molecule, where the feature of x come from, are affected by some uncertainty due to their experimental measure and consequently neither a perfect deterministic model could output an energy value free from uncertainty. Term (23) is known as bias term, it represents instead the error due to the choice of the particular function as predictive model that is different from the true and unknown one. For instance, when a linear model is considered, the bias term represents the error due to the nonlinear component of f that evidently can not be modelled by a linear model. Finally, looking at term (24) one recognises the variance of the model output when x is considered as an input. To sum up, the noise cannot be avoided by any model and it can be changed only acting on the input data, the bias is fixed once a specific model family is adopted and only the variance depends on the model parameters determined by the training on $\mathscr{L}$. Consequently, considering that AL can act only on the labelled test aiming to build the best possible one, it follows that the choice of an AL technique can only affects the variance and thus that the AL technique minimising the generalisation error is the one minimising the variance.

REFERENCES

[1] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015.

[2] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007.

[3] Robert Burbidge, Jem J. Rowland, and Ross D. King. Active learning for regression based on query by committee. In *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, IDEAL'07, pages 209–218, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] W. Cai, Y. Zhang, and J. Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, Dec 2013.

[5] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Demonstrating the transferability and the descriptive power of sketch-map. *Journal of chemical theory and computation*, 9 3:1521–32, 2013.

[6] Brian Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press, Cambridge, 4th ed. edition, 2010.

[7] Valerii Fedorov. *Theory of Optimal Experiments Designs*. 01 1972.

[8] Rong Hu, Sarah Delany, and Brian Mac Namee. Egal: Exploration guided active learning for tcbr. pages 156–170, 07 2010.

[9] Jack O'Neill, Sarah Delany, and Brian MacNamee. *Model-Free and Model-Based Active Learning for Regression*, volume 513, pages 375–386. 01 2017.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016.

[12] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[13] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1070–1079, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[14] J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, 2017.

[15] Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data*, 4:170193–, December 2017.