

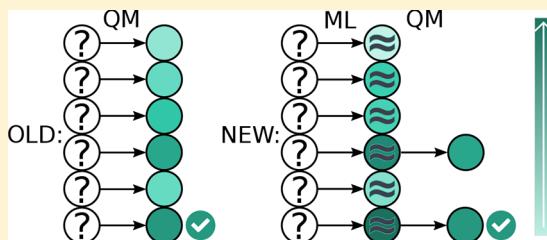
# Acceleration of Inverse Molecular Design by Using Predictive Techniques

Jos L. Teunissen, Frank De Proft, and Freija De Vleeschouwer\*

Research Group of General Chemistry (ALGC), Vrije Universiteit Brussel (VUB) Pleinlaan 2, 1050 Brussels, Belgium

## S Supporting Information

**ABSTRACT:** This study addresses one of the most important drawbacks inherently related to molecular searches in chemical compound space by greedy algorithms such as Best First Search and Genetic Algorithm, i.e., the large computational cost required to optimize one or more quantum-chemical properties. Significant speed-ups are obtained by initial property screening via predictive techniques starting already from very small databases. It is shown that the attainable acceleration depends heavily on the molecular properties, the predictive model, the molecular descriptor, and the current size of the database. We discuss the implementation and performance of predictive techniques in molecular searches based on a fixed molecular framework with a selection of sites to be filled with groups from a chemical fragment library. It is shown that for some properties speed-ups of a factor of 5 to even 20 can be obtained, while inverse design procedures on more complex properties still reach speed-ups of a factor of 2 without losing performance.



## INTRODUCTION

The search and design of new innovative materials displaying specific characteristics are a major aspiration in the field of computational chemistry. Due to the vastness of Chemical Compound Space (CCS), extensive screening is unattainable. Hence, a large variety of computational tools and algorithms have been proposed and employed to search methodically and efficiently for new structures.<sup>1–12</sup> Most of these searches evaluate a number of chemical structures via advanced computational methods.<sup>9,13,14</sup> Paths that lead to well-performing molecules are further explored, while worse performing molecules are neglected. The discard of the less optimal structures is however an unfortunate “waste” of computational effort since they contain useful structure–property correlation. It is therefore advantageous to use this information to speed up the search through chemical space.

In this study, we focus specifically on restricted chemical compound spaces defined by a fixed molecular framework with changeable sites and a library of structural fragments that can be placed on these sites. Finding an optimal structure in these chemical spaces is a combinatorial problem that can be solved by branch-and-bound techniques, Genetic Algorithms, Particle Swarm Optimizations,<sup>15</sup> or Bayesian Optimizations.<sup>16</sup> Although the procedure described in this paper is more generally applicable, we use here, without loss of generality, primarily two algorithms that explore relatively efficiently small chemical spaces: the Genetic Algorithm and the Best First Search algorithm. The Best First Search algorithm (BFS)<sup>17–19</sup> is a branch-and-bound method that searches through a graph (tree) by following the optimal path. From one node, i.e., a specific molecule, all paths toward other nodes are evaluated before selecting the most optimal node as a new starting point.

Normally, the evaluations of the nonoptimal nodes are not used anymore in the searching algorithm. The Genetic Algorithm (GA)<sup>10,20,21</sup> is an evolutionary method that tries to evolve to a population with highest fitness by breeding new generations via crossover and mutations.

The simplicity and robustness of both algorithms make them highly attractive to use. Unfortunately, BFS and GA tend to be quite greedy methods, and since the evaluation of the nodes or the determination of the fitness value of an individual concerns the calculation of a target molecular property, often a computationally expensive operation, requiring one or more quantum-chemical calculations, the final computational cost can become very high.

In the literature, several techniques have been proposed to accelerate and guide the search, usually focusing on large parallel high through-put approaches.<sup>16,20</sup> Kanal et al.<sup>20</sup> used a multistep screening and refinement process to design new organic photovoltaics with multiple desirable characteristics, in which a genetic algorithm was applied as the first step and further filtering steps depended on the computational demand of the properties in question. Hernández-Lobato et al.<sup>16</sup> performed high-throughput virtual screening by collecting hundreds to thousands of parallel data measurements based on a Bayesian optimization combined with Thompson sampling, in which the acceleration is afforded by predicting the most useful data measurements. Additionally, we mention the recently developed PHOENICS optimizer which proves to quickly optimize expensive-to-calculate objectives, although

Received: September 24, 2018

Published: May 7, 2019

mainly focusing on continuous chemical space formulations.<sup>22,23</sup>

In this work, the BFS and GA optimization processes evaluate many chemical structures, resulting in a continuously growing database that is used as input for predictive techniques on the properties to be optimized, as such accelerating the molecular search. When the property is sufficiently predictable and the molecules in the database are sufficiently similar, databases of 50–100 molecules are already enough to provide initial screening, significantly downsizing the number of quantum-chemical calculations. Since in BFS, molecular structures are tuned site by site, the various molecules display a large structural similarity. Equivalently, when mutation rates are small, the structures evaluated in GA also share a large resemblance. Consequently, these databases contain a number of useful quantitative structure–property correlations that can be used to predict property values of new compounds without computing them explicitly.

The preferred predictive technique depends strongly on the nature of the property to predict. Some properties are relatively easy in the sense that each structural element has a certain contribution to the target property that is independent of the remainder of the structure, i.e., the other sites. In these cases, a linear regression model is a good choice. Other properties, however, are largely determined by the presence or absence of certain local features. Hence, the property cannot be expressed as a mathematical equation with every molecular fragment contributing to a certain extent, and nonlinear methods such as machine learning techniques are preferred over statistical linear regression approaches.

The advantage of linear regression is that it requires rather small amounts of data. In addition, the model constructs an explanatory fit that is easily understood in terms of contributions from each model parameter to the property of interest. A drawback of statistical models is that there is no straightforward path to improve the model when more data is available. The ideal number of training samples (e.g., molecular structures and their property values) is preferably somewhat larger than the number of variables (e.g., fragment contributions to the property). Machine Learning approaches, on the other hand, do not require any mathematical model of the underlying system, improve systematically when more training data are present, and can more easily handle multiple parameters with smaller risks of overfitting. Hence, when dealing with a continuously growing training database, in the end, machine learning techniques should always outperform the regression models for larger data sets. The two basic models used in this project are normal ridge regression (NRR) and kernel ridge regression (KRR) as the linear and machine learning models, respectively. The advantage of KRR above other ML methods such as artificial neural networks (ANN) is that training the model is fast and works well with smaller databases as described in this study. When the size of chemical compound space becomes too large for linear or ML models to work well, ANN models are a valuable alternative.<sup>24</sup>

Besides the property, the preferred predictive model depends on the chosen chemical descriptor and the current size of the database. In this study, the universal Bag-of-Bond (BoB) molecular descriptor<sup>3,25</sup> is used alongside a simpler case-specific one-dimensional descriptor where a binary string indicates which group is present on which site (vide infra). The advantage of the binary descriptor is its close relation to how the database is built up, and the linear model can be

understood in terms of contributions from the presence or absence of certain molecular groups. The advantage of the Bag-of-Bonds descriptor is that it contains information about the three-dimensional structure, and hence the model can learn additional contributions from specific molecular orientations.

In this project, we will study how predictive analytics improves searches in chemical compound space performed via inverse molecular design techniques, mainly focusing on the BFS and GA methods. First, we show how predictive analytics as a function of an increasing database depends on the chosen property. Second, the implementation of the algorithm is explained, and the results of some BFS and GA optimization runs are discussed.

## METHODOLOGY

Linear regression via the least-squares method is based on mathematical equations of the form

$$y_i = \sum_{j=1}^N X_{ij}\beta_j, \quad i = 1, 2, \dots, m \text{ and } m > N \quad (1)$$

where  $y_i$  is the property of interest for a structure  $X_i$  that is represented as an  $N$ -dimensional vector.  $\beta$  contains the  $N$  unknown regression coefficients. (When the  $j^{\text{th}}$  dimension is not sampled,  $\beta_j$  is zero.) The problem then reduces to solving the normal equations

$$\beta_{\text{OLS}} = (X^T X)^{-1} X^T y, \quad \beta_{\text{OLS}} \in \mathbb{R}^N \quad (2)$$

in which  $\beta$  represents the coefficient vector of the ordinary least-squares hyperplane, containing all coefficients  $\beta_j$  that minimize the least-squares objective function  $f(\beta)$ :

$$f(\beta) = \sum_{j=1}^m \left| y_i - \sum_{j=1}^n X_{ij}\beta_j \right|^2 \quad (3)$$

Normal ridge regression introduces a hyperparameter  $\lambda$  to the linear regression model which acts as a penalty term to prevent overfitting. The model coefficients are then obtained via

$$\beta_{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T y, \quad \beta_{\text{Ridge}} \in \mathbb{R}^N \quad (4)$$

When the “kernel trick” is applied onto the normal ridge regression model to generalize toward the use of nonlinear functions, a new set of regression coefficients,  $\alpha$ , is determined via

$$\alpha = (K + \lambda I)^{-1} y, \quad \alpha \in \mathbb{R}^n \quad (5)$$

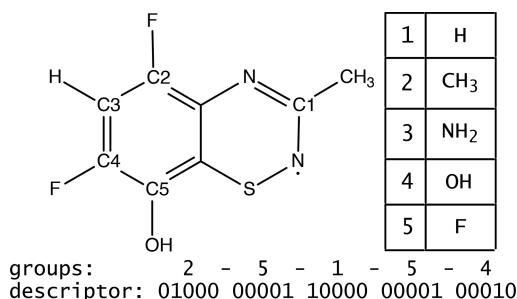
where  $n$  is the number of training samples, and the Gaussian kernel matrix,  $K$ , contains the inner products of the input samples  $X$ <sup>26,27</sup>

$$K_{ij} = \exp \left[ -\frac{\|X_i - X_j\|^2}{2\sigma^2} \right] \quad (6)$$

where  $\|X_i - X_j\|$  stands for the Euclidean norm or distance.  $K_{ij}$  is therefore a quantitative measure of similarity between a new molecule  $i$  and training molecule  $j$ . The  $\sigma$  parameter is an additional hyperparameter that represents the length scale of the kernel. The  $\lambda$  and  $\sigma$  hyperparameters were optimized using 3-fold cross-validation technique on a logarithmic parameter grid.

To predict a property of a molecule one also needs to choose a molecular descriptor that converts the information encoded in a 3D molecular structure to an appropriate vector of numbers. In accordance with previous studies, we use the Bag-of-Bonds (BoB) representation,<sup>28</sup> since it outperforms the Coulomb matrix descriptor<sup>29</sup> in all our cases. The Coulomb matrix representation<sup>10</sup> makes use of the so-called Coulomb matrix C, using atomic coordinates  $R_i$  and nuclear charges  $Z_i$ . The main diagonal of the Coulomb matrix consists of a polynomial fit of the nuclear charges to the total energies of the free atoms, while the remaining elements contain the Coulomb repulsion for each pair of nuclei in the molecule. In the Bag of Bonds representation all entries in the Coulomb matrix are put in different bags corresponding with their bond type (C–H, C–O, C–C, O–H, . . .).

In the BFS and GA algorithms, a compound is represented as an array of numbers, indicating the functional group, with each array position corresponding to a position in the molecular framework. The databases consist of specific configurations of functional groups on these sites and their assigned property values. The most straightforward way to define the structure–property relations is to relate the placement of each functional group on a certain site to a particular property contribution. In fact, the BFS algorithm principle relies on this independence of sites for convergence. Accordingly, the simplest possible molecular descriptor specifically indicates each functional group on each site. We will coin this descriptor a one-dimensional array of functional groups, abbreviated as 1DL. In 1DL, a molecule is simply described as a one-dimensional array of sites indicating per site - via an integer - the functional group that is present. Any structure X is encoded as the vector  $(b_1, b_2, \dots, b_i, \dots, b_N)$ , with  $b_i$  representing the functionalization on site  $i$ . This array is then converted to a binary flattened list where every number  $b_i$  is represented as a list of zeros except for a one at the position indicating the functional group present on that site. An example is shown in Figure 1.



**Figure 1.** Example of the one-dimensional descriptor (1DL). In this case, a random thiadiazinyl derivative with 5 modifiable sites using a library of 5 fragments.

This work is based on two different data sets. Both databases were obtained by searching for optimal property values (minimization or maximization) via the BFS method in a chemical space defined by a molecular framework having several adjustable sites and a library of chemical fragments that can be placed on these sites. Note, however, that since the databases were built by searching for optimal properties, our data sets will be biased, and extremes therefore occur more frequently.

Data set 1 contains 4794 HOMO and LUMO energy values from a chemical space defined by the adamantane framework (10 sites) and a library with 15 different molecular fragments.<sup>30</sup> Their frontier-orbital energy values belong to molecular geometries obtained by the B3PW91<sup>31</sup> functional and the 6-31G(d,p)<sup>32</sup> basis set as implemented in the Gaussian09<sup>33</sup> software package. Data set 2 contains the electrophilicity index and radical stability values of 5434 thiadiazinyl radical derivatives.<sup>34,35</sup> These derivatives come from a chemical space defined by the thiadiazinyl radical framework with 5 adjustable sites and 21 possible fragments. The electrophilicity index  $\omega$  can be calculated as<sup>36</sup>

$$\omega \approx \frac{(I + A)^2}{8(I - A)} \quad (7)$$

with  $I$  representing the vertical ionization potential, and  $A$  representing the vertical electron affinity (in eV). The radical stability  $\text{stab}_X$  of a thiadiazinyl radical derivative  $X$  having a nitrogen atom as the radical center can be computed via the following model<sup>37</sup>

$$\text{stab}_X = \text{BDE}(X - H) - \text{stab}_H - a\Delta\omega_X\Delta\omega_H - b\Delta\chi_X\Delta\chi_H \quad (8)$$

where BDE is the bond dissociation enthalpy of  $X - H$ ,  $\Delta\omega_X = \omega_X - 2$  is the electrophilicity index difference with 2 eV, the estimated boundary between electrophilic and nucleophilic radicals, and  $\Delta\chi_X = \chi_X - 3$  is the Pauling electronegativity difference with 3, the estimated boundary between strongly and weakly electronegative atoms, for the atom of  $X$  directly involved in the bond with H. The model reduces to the following equation (in kJ/mol):

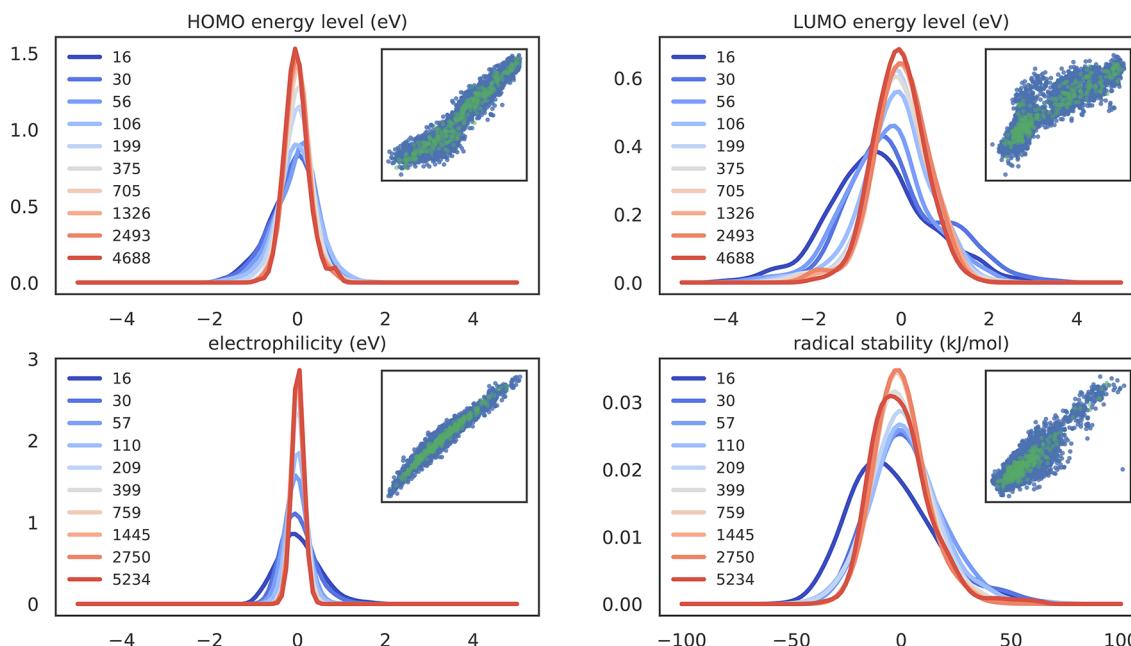
$$\text{stab}_X = \text{BDE}(X - H) - 235.7 + 0.7997\Delta\omega_X - 8.7 \quad (9)$$

Geometries were optimized using B3LYP/6-31G\*.<sup>32,38,39</sup> The electrophilicity index was computed with B3LYP/6-311+G\*\*. Bond dissociation enthalpies were obtained using the B3P86/6-311+G\*\*<sup>32,40</sup> level of theory.

## RESULTS AND DISCUSSION

**Global or Local Properties?** First, we assess how the different properties behave using the concept of locality. When the property of a molecule is local, we mean that its property value is determined by only a confined topological part. For example, the dipole moment of butanol is most significantly determined by the hydroxyl group as opposed to the butyl chain. In contrast, for a global property every part of the structure contributes to a significant degree. For example, every atom in a molecule contributes to the molecule's total volume. In simple cases, a global property can be expressed as a sum of contributions from each structural element. When cross-terms are neglected, i.e., when the contribution of one structural element does not depend significantly on neighboring sites, all sites can be considered independent. An example of such an independent property is the molecular weight. The molecular weight of a structure is simply the sum of the mass of the individual elements. When considering a molecular scaffold consisting of several sites containing different chemical fragments, this is called the independent site approximation (ISA), in which the property contribution of a structural element on one site is independent from the property contribution of another structural element on a different site.

## Linear regression for increasing database sizes.



**Figure 2.** Kernel Density Estimation plots for the linear model with the 1DL descriptor applied on different properties and using increasing database sizes. The insets show predicted vs real values for 300 training samples (green) and the rest of the database as test samples (blue).

In this case, the property  $P$  for any structure  $X$  can be written as

$$P(X) = P(X_{\text{framework}}) + \sum_{i=1}^{N_{\text{sites}}} P(X_{\text{group},i}) \quad (10)$$

This expression will be almost but not entirely valid for a property such as the electrophilicity index  $\omega$ . One can after all imagine that the contribution of an electron withdrawing group is enhanced when a strongly electronegative group is placed on a neighboring site. Nevertheless, it turns out that good structural searches and predictions can be made when assuming that the ISA is valid for this property.

Two other properties we will examine in this study, i.e., the frontier-orbital energies, are trivially very dependent on to what degree the orbitals are extended over the structure. When the molecular orbital is fully delocalized, it has contributions from atomic orbitals on every atom. In other words, every structural fragment contributes in some way to the total molecule's orbital energy level. In such cases, the ISA will be largely valid, and the linear model will perform well. However, when the molecular orbital is rather localized, e.g., on a functional group, large parts of the molecule are not relevant for the orbital energy level. Hence, the orbital energy will behave local and nonlinear, and therefore the independent site approximation does not apply.

To test whether properties behave locally or globally, it is checked to what degree the ISA is fulfilled. When the ISA is fulfilled, a linear regression model with the 1DL descriptor should result in a perfect correlation. Figure 2 shows the results of the linear regression model with the 1DL descriptor on the four different properties (HOMO and LUMO energy from data set 1 and electrophilicity index and radical stability from data set 2) for increasing training sizes.

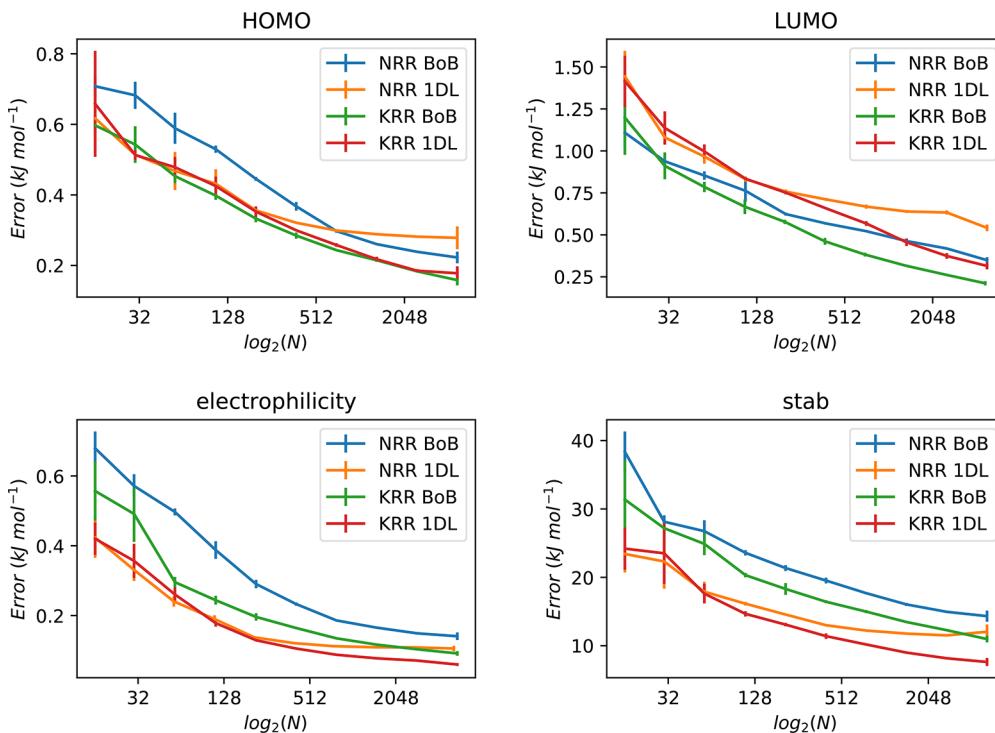
From Figure 2 one can observe that both the HOMO energy level and the electrophilicity index are well predictable via the

linear model, as shown by the width of the probability density peaks and correlation diagrams. In many adamantane derivatives the HOMO is delocalized over the whole cage. This means that replacing any functional group in an adamantane derivative affects the HOMO energy level, so the orbital energy level can be largely understood as a sum of contributions from each functional group plus the contribution from the molecular framework.

In contrast to the HOMO, the LUMO in the data set of adamantane derivatives is often localized on one specific part of the molecule. Note that this is partly due to the type of optimization that was applied, i.e., the data set contains mainly structures related to a minimization of the LUMO energy level. For example, when considering an adamantane derivative containing an electron withdrawing group such as  $\text{NO}_2$ , the LUMO is mainly localized on the  $\pi^*$ -orbital of that functional group. In this case, other parts of the molecule contribute to a much lesser extent to the energy level of the LUMO. Hence, the LUMO energy in our database is largely a local property for which the ISA is invalid, and linear predictions via the list-of-sites descriptor are poor.

The electrophilicity index  $\omega$  can be computed from the vertical ionization potential,  $I$ , and the vertical electron affinity,  $A$  (see eq 7). As the electron affinity is always considerably smaller than the ionization potential,<sup>41</sup>  $\omega$  is mostly determined by  $I$ . Since, via the Koopmans' theorem,  $I \approx -E_{\text{HOMO}}$ , and the HOMO is also in the case of thiadiazinyl derivatives largely delocalized over the whole molecule, the electrophilicity index will behave as a global property, for which the ISA is largely valid.

Lastly, the radical stability values are considerably harder to predict. The radical stability is a function of both the homolytic bond dissociation energy (BDE) with hydrogen and the electrophilicity index.<sup>37</sup> Unlike the electrophilicity index term, the BDE cannot be understood in terms of the frontier molecular orbitals only.<sup>42</sup> Moreover, the model does not



**Figure 3.** Plots of root-mean-square errors (RMSE) vs database size for both the normal ridge regression and the kernel ridge regression methods using the 1DL and BoB molecular descriptors on four different properties.

**Table 1.** Average  $R^2$  Values of the Predictions for the Test Set, Using Normal Ridge Regression (NRR) and Kernel Ridge Regression (KRR) as Prediction Models and Using Either the List-of-Sites (1DL) or the Bag-of-Bonds (BoB) Descriptors Fitted on 300 Training Samples<sup>a</sup>

	property	1DL		BoB	
		NRR	KRR	NRR	KRR
data set 1	HOMO	0.956 (0.33)	0.959 (0.31)	0.913 (0.33)	0.963 (0.30)
	LUMO	0.893 (0.72)	0.907 (0.68)	0.935 (0.61)	0.955 (0.47)
	HL-gap	0.903 (0.80)	0.913 (0.76)	0.938 (0.65)	0.951 (0.58)
data set 2	$\omega$	0.977 (0.13)	0.981 (0.12)	0.904 (0.27)	0.960 (0.17)
	stab	0.864 (14.0)	0.897 (11.8)	0.720 (19.0)	0.780 (17.0)

<sup>a</sup>Root-mean-square errors are given within brackets with units eV or kJ/mol (stab).

improve any further when increasing the training set from 2750 to 5234 training samples.

**Performance of Nonlinear Models.** Since the LUMO energy level and stab (via the BDE) behave as local properties and cannot be predicted properly using linear models, nonlinear models such as kernel ridge regression need to be applied. For each property, the performance of the model (NRR and KRR) and descriptor (1DL and BoB) combination was evaluated for increasing database sizes as displayed in Figure 3, whereas correlation coefficients and root-mean-square errors (RMSE) for the case of 300 training samples are given in Table 1.

From the correlation coefficients one can deduce that the electrophilicity index and the HOMO energy level are well predictable. For the 1DL descriptor, the linear and nonlinear models perform equally well. The predictions result in outstanding RMSE values of approximately 0.3 eV for the HOMO energy and 0.15 eV for  $\omega$ . The more locally behaving LUMO energy level is harder to predict and therefore also the HOMO–LUMO gap. The correlation coefficients obtained with the KRR model in combination with the BoB descriptor

are very reasonable, but the RMSE values are twice as high as for the HOMO energy. The predictions for the radical stabilities are further off, especially when the BoB descriptor is used.

As can be observed from Figure 3 and Table 1, the performance of the kernel ridge method is always similar to or better than the linear model, and its improvement becomes more pronounced when going to larger training set sizes. This is expected since this model contains many more parameters to fit when the training sets are larger than a few hundred samples. The BoB descriptor performs well in combination with the kernel regression but worse with the normal regression since the individual entries of the BoB are not directly related to specific functional groups, so there is no reason to assume a linear relationship between the BoB values and the properties when using normal regression.

The predictions made via kernel regressions using the 1DL descriptor are adequate regarding that the descriptor contains only 10 or 5 integers, i.e., the number of adjustable sites in the molecular framework for adamantane and thiadiazinyl, respectively. Especially when the database is small, this

generally outperforms the more complicated kernel-BoB model; only for the LUMO the BoB descriptor clearly performs better than the 1DL descriptor at small database sizes. Not surprisingly, the kernel regressions outperform the linear regression except for very small database sizes. Since the number of parameters in kernel regression is the same as the number of training samples, the model can continuously improve when the database is enlarged. The linear regression model, on the other hand, contains a fixed number of parameters and hence from a certain amount of training data a plateau is reached, and no significant improvements can be made anymore. In the prediction of the electrophilicity index, also the kernel regression reaches a plateau; however, this occurs because no better predictions can be made than the error made in the data itself, and therefore the improvement rate stagnates.

**Implementation in the BFS and GA Procedures.** The best prediction model depends on the property and the database size. On top of that, predictions based on a large-dimensional descriptor, such as the BoB, can be underfitted when few training data is available while outperforming simpler descriptors for larger database sizes. Hence, the implementation of predictions into molecular search methods should consider the continuous growth of the database. We therefore decided to use simultaneously multiple prediction models in our chemical design algorithm, which are repeatedly retrained.

In every (re)training step the current database is split into a validation set and a set for hyperparameter optimization. The size of the hyperparameter optimization set is either 75% of the total database or maximally 300 samples. The hyperparameter set is used for a grid search hyperparameter optimization involving a 5-fold cross-validation scheme as implemented in the scikit-learn package.<sup>43</sup> The model showing the highest correlation of determination,  $R^2$ , on the validation set is then selected as the predictive model. All hyperparameters are tuned only at the beginning of each global iteration, i.e., one full run over all sites in the BFS procedure or one generation in the GA approach. The model hyperparameters that presented the best results are then used for all other predictions in that global iteration.

In the BFS and GA approach, we usually consider several structures simultaneously, for which we analyze the properties before constructing a new set of structures. In BFS, this set consists of structures that differ only in a single functionalization, i.e., on a particular site all possible functionalizations from our fragment library are evaluated while keeping the other sites fixed. In analogy with the Genetic Algorithm, we will call this set of structures a population. The population size is generally in the order of 10 to 50 individual molecular structures.

One now needs to decide whether the property of a molecular structure needs to be calculated computationally based on the overall quality of the predictions. When predictions are far off from their computed values, we fall back on computations via quantum chemistry software; however, when the predictions are rated as quantitatively good, we use the predicted values from the best prediction model for a certain fraction of the population. As a starting point, we decided to use a simple heuristic function that decides which fraction of the *ab initio* calculations needs to be performed. Using a fraction saves the trouble of choosing a cutoff value. The fraction depends on the  $R^2$  value of the best predictive model because it is independent of the property values unlike the RMSE and MAE, and therefore the same

heuristic function can be used for any given property. Moreover, the values in Table 1 indicate that for the examined properties  $R^2$  and RMSE correlate almost perfectly. When  $R^2$  is 0.75, it means that 75% of the variance in the property is explained by the model. From this point, we regard the model as good enough to be used. By choosing the cutoff as 0.75 we made a balance between the reduction of computer time and the probability of overlooking good candidates. When  $R^2$  is 1.0, the model exactly represents the system, and, in principle, no explicit calculations need to be performed; in practice we incorporate an extra check by performing an *ab initio* calculation on the best structure as predicted by the most suitable model. Hence, the fraction of predictions is increased from 0.0 to 1.0 when the  $R^2$  value goes from 0.75 to 1.00 as given by the formula

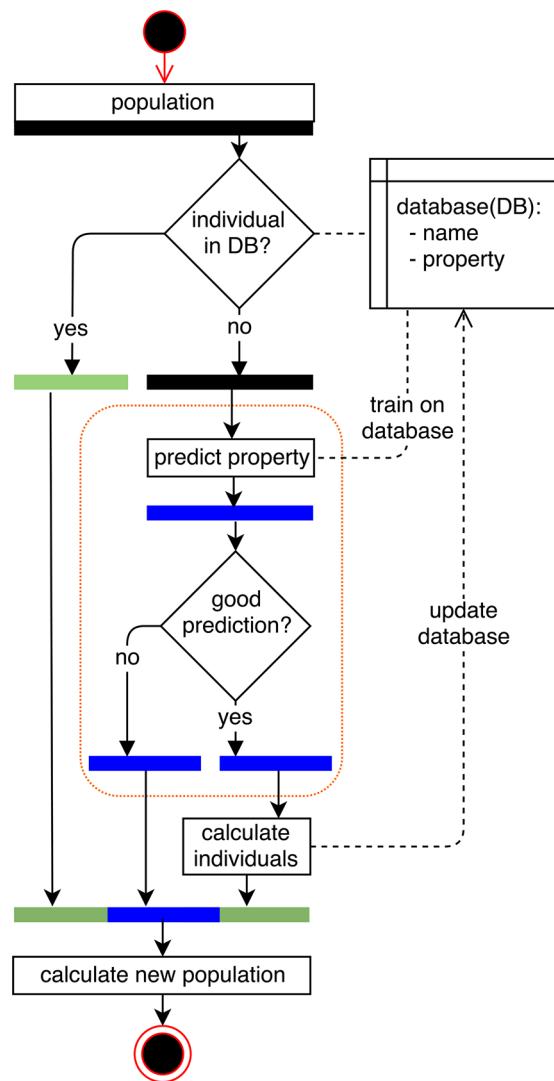
$$f(R^2) = \min(1, 4 - 4R^2) \quad \text{for } R^2 \text{ in the range [0, 1]} \quad (11)$$

In the Genetic Algorithm formulation, the predictions are applied when the database contains at least 50 samples. For the BFS optimization procedure, predictive techniques are only applied when one global iteration has finished. Thence, for each functional group on each site at least one molecule is present in the database, and consequently there exists at least one data-point for each parameter in the linear regression model with the 1DL descriptor.

The workflow for one optimization step is depicted in Figure 4. First, a population is constructed. Next, it is checked whether structures of this population are already present in the database. Once the database is large enough, the property values of the individuals that are not yet in the database are predicted. When the  $R^2$  value of the best predictive model is higher than 0.75, the population is split into two sets with their fraction based on the heuristic function defined in eq 11. We have chosen to calculate the property values of the most promising individuals with the computationally more expensive quantum chemistry software method. Finally, the computational results are added to the database, and a new population is constructed based on the structure with the best property value.

This workflow is easily adaptable to multiple property optimizations, where the best predictive model is taken as the model with the best overall performance (over all properties). On top of that, in the kernel ridge regression model, the same kernel can be used for multiple properties, hence saving the computationally demanding step of inverting the kernel matrix.<sup>44</sup>

**Testing the New Algorithms.** To test the new algorithms, several property optimizations were run with and without application of the predictive models. The same molecular framework and fragment library were used as in the construction of the initial databases. However, the models only tap from the database that is created during the optimization itself. Comparing a run with and without predictive component is not straightforward. When the prediction of an unknown structure, to be imagined as the possible optimum, is too far off, it might not be selected to be calculated with the quantum chemistry program. Hence, the poorly predicted value will not be corrected, and the algorithm will continue with another structure as the new starting point. As a result, both runs will diverge. Nevertheless, we can compare the number of *ab initio* calculations needed per global



**Figure 4.** Workflow for the optimization of one site in the BFS algorithm, in which at least one of the predictive models can be used. The black bars indicate a set of individuals with no property data. The green bars indicate that the individuals have property values calculated via the ab initio method. The blue bars indicate a set of individuals with only predicted property values.

iteration starting from the same initial starting configuration and sequence in which the various sites are visited.

For each property, three BFS optimizations were performed, two of them starting from the unsubstituted molecule (i.e., adamantane or thiadiazinyl) and one from a randomly generated molecular derivative. In addition, the site sequence was varied, from “normal”, i.e., following the site order consecutively from 1, 2, . . . , to  $n$ , to completely random. The BFS was terminated when the optimum structure could not be improved anymore by changing a single site. Note, however, that the optimal structures are often local optima. In a previous work,<sup>30</sup> we adapted the gradient-directed Monte Carlo approach as applied by Hu, Beratan, and Yang<sup>45</sup> to the discrete BFS algorithm. To escape from possible local traps, after each complete BFS convergence, a random structure is generated and accepted with a certain probability using the Metropolis algorithm and then used as the new starting point for the next BFS cycle. For more details on the procedure, see ref 25. For two property optimizations, additional Monte Carlo

steps on the BFS optima were applied (gap-MC and stab-MC) to get an idea of how the predictions improve when the database size increases.

The Genetic Algorithm runs show clear performance enhancement as well. The mutation and crossover probability were 0.02 and 0.90, respectively, and a roulette wheel selection procedure was employed.<sup>46</sup> The GAs ran for 20 generations with a population size of 20 individuals, such that the number of calculations remains similar to the BFS procedure.

In Table 2, we observe that in most cases the BFS-ML will find the same or nearly the same optimum structure as the nonadapted BFS runs. Only larger discrepancies between the two procedures are marked down for the minimization of the HOMO–LUMO energy gap, for which the BFS-ML is performing worse. This can be explained by the chemical space defined by the HOMO–LUMO energy gap of the adamantane derivatives having a number of local minima, which causes different BFS runs, and therefore also BFS-ML runs, to converge to different optima. This could be remedied by taking Monte Carlo steps, as discussed in our previous work.<sup>30</sup> Similar chemical space complexity is seen for the minimization of the LUMO energy, although in this case the differences between the adapted and nonadapted procedures remain small.

In general, the predictability of new structures is excellent with regression coefficients mostly around 0.9 or higher. The performance of the KRR-1DL, KRR-BoB, and NRR-1DL models is very similar, but the most sophisticated model, KRR-BoB, works best in most cases. The high  $R^2$  values are not surprising, considering the high similarity between structures in the constructed databases. Only the prediction of the radical stability values appears to be more difficult, and regression coefficients attain only around 0.7–0.8. These are borderline cases for the heuristic function that was used; therefore, the obtained speed-up is meager. For well-predictable properties such as the electrophilicity index, the speed-up varies from a factor 2 to even 10–20. Excluding the first global iteration step which is needed as the initial database for the predictive models, the fraction of structures requiring quantum-chemical calculations can thus be reduced to a mere 5% of the database size. Note also the 10-fold speed-ups reached for the LUMO minimization, a result that was not entirely anticipated given the quite large root-mean-square errors for smaller database sizes in Figure 3.

The data in Table 2 indicate that attainable speed-ups are highly dependent on both the property and the type of optimization (min versus max). For instance, predictions for the maximization of the electrophilicity index are very successful as reflected by the nearly perfect correlation coefficients and vast speed-ups, whereas for the minimization of the same property the enhancement is undeniably less outspoken. The difference in success can in this particular case be traced back to the dissimilar substituent effect on the thiadiazinyl radical, resulting in highly electrophilic and only moderately nucleophilic radical derivatives for the maximization and minimization of  $\omega$ , respectively.<sup>28</sup> There is also no correlation between the machine-learning performance for a certain property and whether that property fulfills the independent-site approximation which is inherently related to the BFS approach, as can be witnessed from the variation in speed-ups between stab and LUMO energy (ISA not valid). Moreover, note that ISA validity does not guarantee that the NRR-1DL model will be selected. Both for the electrophilicity

**Table 2. Results from All BFS Runs with and without Machine Learning for Three Cases: 1) Unsubstituted Framework and Specific Site Sequence (Site 1, 2, . . . , n); 2) Unsubstituted Framework and Random Site Sequence; and 3) Random Start Structure and Site Sequence<sup>c</sup>**

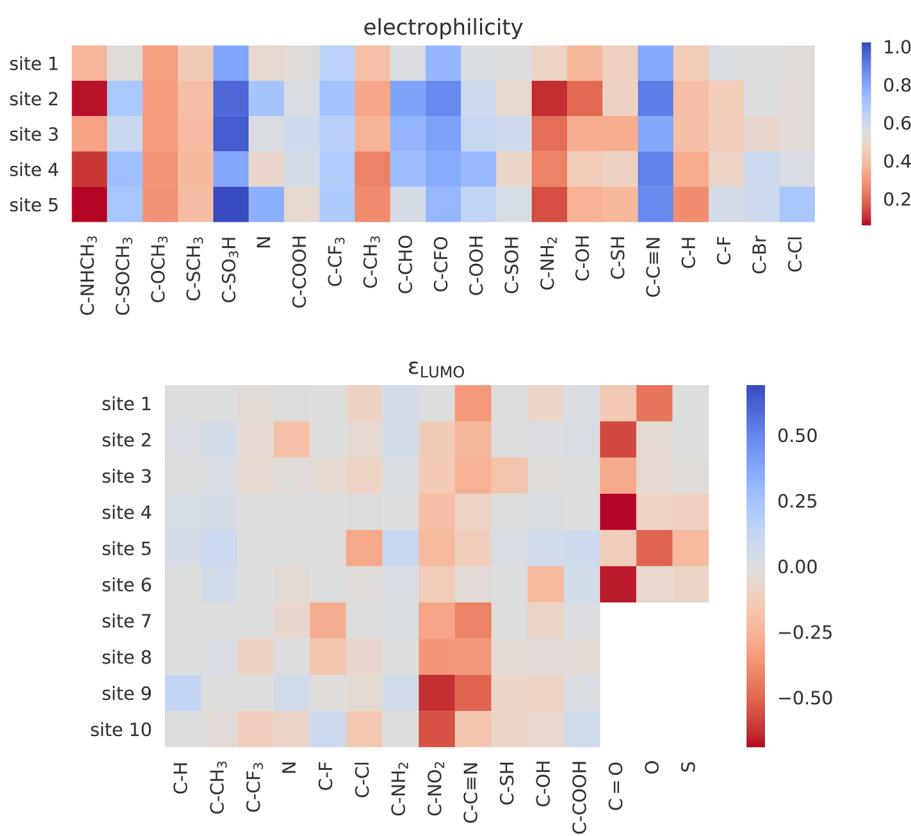
property	start X	site sequence	min/max	no ML				ML				improvement factor
				nGI	ncalc	opt value	nGI	ntotal	ncalc	opt value	N/K	
gap	nosubs	normal	min	3	343	2.91	2	243	152	3.05	K	B
	nosubs	random	min	3	374	3.12	4	385	178	3.58	KNN	B11
	random	random	min	5	560	2.26	4	485	268	2.81	NNK	111
	nosubs	normal	max	2	162	10.81	2	162	132	10.81	K	B
	nosubs	random	max	3	257	10.78	3	257	149	10.78	K	B
	random	random	min	6	644	10.63	4	398	198	10.45	K	B1B
HOMO	nosubs	normal	max	2	276	-4.70	2	276	161	-4.70	K	B
	nosubs	random	random	3	293	-4.67	4	357	187	-4.67	K	B
	random	random	min	4	384	-4.65	3	373	155	-4.67	K	B
LUMO	nosubs	normal	min	5	513	-5.18	6	662	179	-4.93	K	B11B
	nosubs	random	min	3	318	-4.38	4	384	159	-4.38	NNK	B11
	random	random	min	4	448	-4.90	4	413	158	-4.97	KKN	B11
electrophilicity index	nosubs	normal	min	4	321	1.23	3	261	170	1.26	K	B
	nosubs	random	min	2	221	1.23	3	221	153	1.23	K	B
	random	random	min	3	321	1.23	2	181	113	1.23	K	B
	nosubs	normal	max	2	181	5.07	2	181	105	5.07	K	B
	nosubs	random	min	3	200	4.99	2	181	105	4.91	K	B
	random	random	min	2	181	5.07	2	181	109	5.07	N	1
stab	nosubs	normal	min	4	299	11.0	4	299	284	11.0	N	1
	nosubs	random	min	2	181	17.5	2	181	149	17.5	N	1
	random	random	min	2	181	11.6	2	181	181	11.6	N	1
gap - MC	nosubs	normal	min	13	1511	2.91	17	1834	411	2.97	K <sup>a</sup>	B <sup>a</sup>
stab - MC	nosubs	normal	min	14	1039	11.0	13	658	500	11.0	K <sup>a</sup>	1 <sup>a</sup>

<sup>a</sup>Overall best predictive model. <sup>b</sup>Average R<sup>2</sup> value taking best predictive model at each step. <sup>c</sup>The first global iteration is excluded from the improvement values. The best model per global iteration is given with multiple entries when different global iterations had different optimal predictive models. nGI = number of global iterations, nc alc = number of molecules effectively calculated, ntotal = total number of molecules that are either calculated or predicted, N = normal ridge regression, K = kernel ridge regression, 1 = 1DL, B = BoB, MC = Monte Carlo step.

**Table 3.** Results from All GA Runs and Two Bayesian Optimizations (Gaussian Process, GP) with and without Machine Learning<sup>a</sup>

property	min/max	no ML			ML			improvement		
		nGen	ncalcs	opt value	ntotal	ncalcs	opt value	R of best prediction model	fraction	factor
gap	min	20	223	3.87	223	175	3.62		78%	1.4
gap	max	20	294	9.31	226	134	9.33		59%	2.1
HOMO	max	20	264	-5.14	280	188	-4.98		67%	1.7
LUMO	min	20	300	-4.09	264	159	-3.77		60%	2.0
electrophilicity index	min	20	178	1.53	122	90	1.53		74%	1.8
electrophilicity index	max	20	120	4.74	121	79	4.63		65%	2.4
electrophilicity index high mutation rate of 0.2	max	20	406	4.91	411	199	4.97		48%	2.4
Bayesian optimization (GP): electrophilicity index	min	20	210	1.23	205	87	1.31		24%	4.2
Bayesian optimization (GP): LUMO	min	20	206	-3.99	210	143	-3.99		58%	1.7

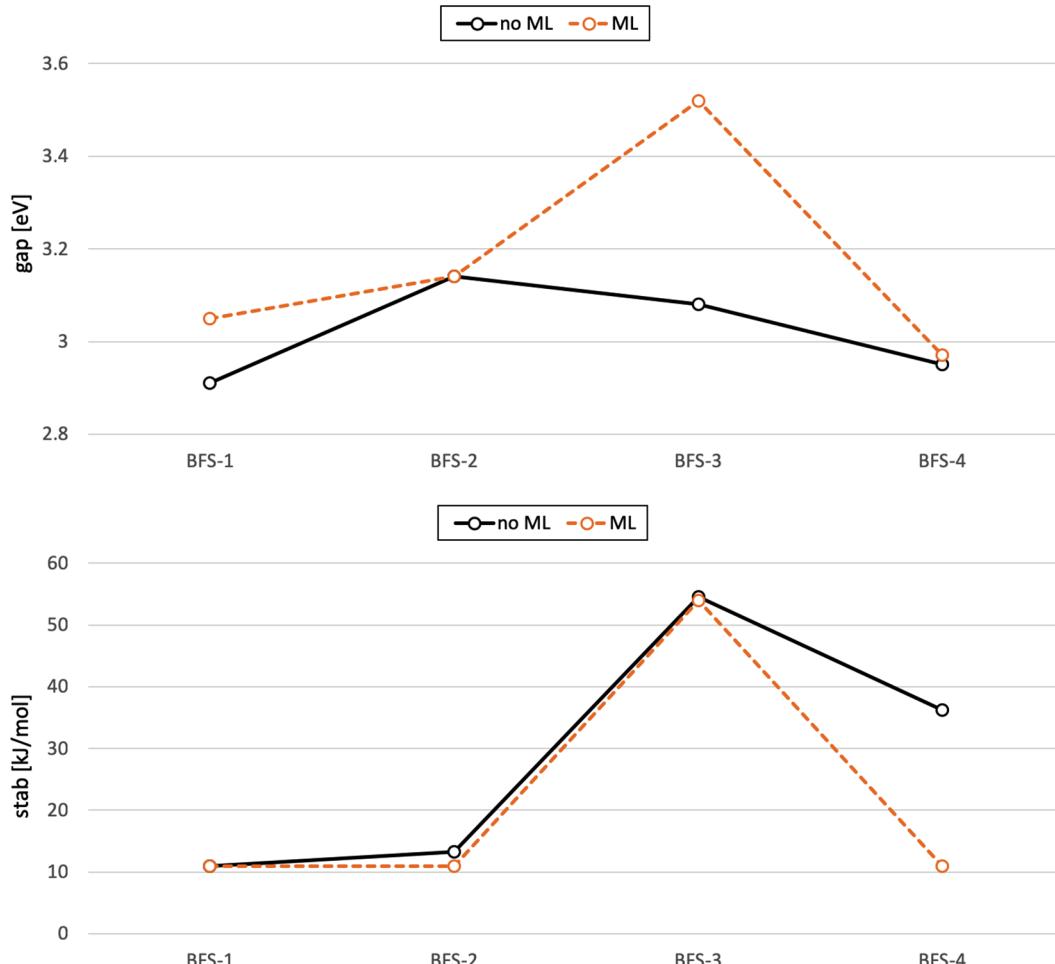
<sup>a</sup>nGen = number of generations, ncalc = number of molecules effectively calculated, ntotal = total number of molecules that are either calculated or predicted. The Pearson's correlation coefficients R for every generation are given in figure format in which the color of the marking represents the best predictive model. Blue = NRR-BoB, orange = NRR-IDL, green = KRR-BoB, and red = KRR-IDL.



**Figure 5.** Heatmaps of the last trained linear model with the 1DL descriptor for the electrophilicity index maximization and LUMO minimization BFS runs. Coefficient units are in eV.

index and the HOMO energy, for which the ISA is largely valid, the KRR-BoB combination is preferred over the NRR-1DL model, even though differences might be negligible, whereas for stab the NRR-1DL model was favored.

The results of GA, as listed in Table 3, with and without predictive component are very similar while the computational cost is reduced by a factor 2. Since the structural diversity within a Genetic Algorithm population is higher than in BFS,



**Figure 6.** Optimal values of the HOMO–LUMO energy gap and the radical stability after 4 BFS runs connected via Monte Carlo steps, with and without predictive methods (no ML vs ML).

the predictive model needs a larger database to adequately predict the property; hence, the correlation of small databases is smaller. Except for the minimal HOMO–LUMO gap, we clearly detect an improvement of the predictive models throughout the optimization process. We also observe that the best performing model changes quite often, indicating that any of the model–descriptor combinations can be suitable for GA.

To illustrate that this procedure is more generally applicable, we performed additional Bayesian optimizations. These optimizations require that the acquisition function can sample multiple points simultaneously.<sup>47</sup> The Bayesian optimizations were performed by modeling the surrogate property landscape as a Gaussian process in a space of only categorical dimensions. The optimization starts by evaluating 10 random points, and, subsequently, for every generation 10 new molecules were sampled and treated as a new population to use the implementation as shown in Figure 4. Two optimizations were performed, one minimizing the electrophilicity index and one minimizing the LUMO energy. For both Bayesian optimizations, good results were obtained, and the number of calculations was reduced significantly by use of predictive analytics indicating that the procedure as described in this paper is also applicable to other search algorithms.

The linear regression models perform almost equally well as the kernel regression methods. Additionally, they have the

advantage that these models can be investigated to see the average contributions of each functional group to the property for the current database. This is demonstrated for the two last trained linear models with the 1DL descriptor for the LUMO orbital energy minimization and the electrophilicity index maximization (see Table 2). In Figure 5, we depict the contributions of each functional group to the property of interest for the current data set. Hence, the figures reflect predominantly structures similar to the current optimum. It can be observed that the coefficients for the LUMO model are very site dependent, while in the case of the electrophilicity index, coefficients have similar values for the different sites. The site-dependence of the LUMO orbital energy coefficients indicates that currently the LUMO is mostly minimized by carbonyl and nitro groups on very particular sites, with only very small contributions from functional groups on other sites. In other words, the LUMO is mainly residing on the nitro or carbonyl groups, indicating that the LUMO energy is a local property. In contrast, when considering the electrophilicity index, all sites can significantly contribute, with the type of functionalizations as the most determining factor, largely independent of which sites these groups are located. Moreover, the heatmap already provides information on which functional groups might minimize the electrophilicity index. In this case, we are clearly dealing with a global property.

Since all the optimizations started without any database at forehand, still a considerable number of calculations had to be performed. However, in practice, one performs several runs, e.g., in combination with Monte Carlo steps, and one can therefore start from an initial database. This should significantly reduce the number of computationally cost-demanding calculations. Therefore, we have performed for the first entry of the gap and stab minimization in **Table 2** (start X = nosubs, site sequence = normal), which are among the least accelerated property optimizations, three additional BFS runs connected via Monte Carlo steps. First of all, we note that the additional BFS runs for these particular cases do not (significantly) improve the optimal gap and stab values in entry 1, indicating that the first BFS minimization was already quite effective. The optimal values after each BFS run are presented in **Figure 6**. Nonetheless, especially in the case of the radical stability a notably reduced computational workload is observed with respect to the single BFS run, as is shown by the fraction of computed vs total number of structures that were visited during the optimization process (but excluding the first global iteration), i.e., the workload decreases from 92% after BFS-1 to 72% after BFS-4. Regarding the gap minimization, the increase in acceleration rate seems less pronounced. However, when taking into account the actual workload by including the first global iteration in the improvement values, one observes an almost 3-fold rise in acceleration rate from 1.6 to 4.5. Moreover, the prediction models show improved  $R^2$  values compared to BFS-1. In **Table 2**, only the average  $R^2$  value, taking the best predictive model at each step, is mentioned. Remarkably, the further along the BFS process the more a particular prediction model stands out as the best, namely the kernel ridge regression with the Bag-of-Bonds descriptor for the gap minimization and KRR with the 1DL descriptor for stab. Kernel ridge regression outperforms normal ridge regression because for the former the number of variables scale with the size of the database.

Given these results, it becomes now feasible to tackle larger design problems and to assess vaster chemical spaces, using not only rather transparent and relatively easy-to-implement algorithms such as BFS and GA but also more advanced chemical search algorithms such as Bayesian optimizations.

## CONCLUSION

In this study, we have exploited predictive techniques on quantum chemical properties, based on databases constructed on-the-fly, to accelerate molecular searches in chemical compound space mainly via the Best First Search and Genetic Algorithm approach, but also two Bayesian optimizations were included. In this paper, we specifically focused on fine-tuning known molecular scaffolds, for which we anticipate finding the best results. We found good to excellent performance for most target properties, independent from their global or more local character and whether they fulfill the independent-site approximation inherently related to BFS. For GA and Bayesian optimizations, the registered speed-ups were in general smaller than for BFS, which can be related to the structural diversity in their respective databases. In addition, it was illustrated that for different target properties different prediction techniques may perform best, either in terms of the molecular descriptor or the algorithm itself. Moreover, we have shown that significant speed-ups (up to 20 times) can be obtained using the suggested heuristic function which, based on the performance of the predictive model, decides the fraction of the total

population requiring quantum-chemical calculations and the fraction that can be predicted. Finally, it was illustrated for two cases that, when the linear regression model performs well, relevant structure–property relations can be derived in terms of functional group preference and site-dependency. This work opens up the possibility for faster and broader molecular searches allowing us to find new and improved structures for complex and challenging molecular optimization problems.

## ASSOCIATED CONTENT

### S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acs.jcim.8b00654](https://doi.org/10.1021/acs.jcim.8b00654).

Two data sets: data set1.json, 4794 adamantane derivatives with HOMO and LUMO energy values; data set2.json, 5434 thiadiazinyl radical derivatives with electrophilicity index and radical stability values (ZIP)

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [fdevlees@vub.be](mailto:fdevlees@vub.be).

### ORCID

Freija De Vleeschouwer: [0000-0003-0563-1509](https://orcid.org/0000-0003-0563-1509)

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

Computational resources and services were provided by the shared ICT Services Centre funded by the Vrije Universiteit Brussel, the Flemish Supercomputer Centre (VSC), and the FWO. F.D.P. acknowledges the Fund for Scientific Research-Flanders (FWO) and together with J.L.T. and F.D.V. the Free University of Brussels (VUB) for continuous support to his research group, in particular the VUB for awarding a Strategic Research Program (SRP) to the ALGC research group started on January 1, 2013. F.D.P. also wishes to acknowledge the Francqui foundation for a position as Francqui research professor.

## REFERENCES

- (1) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.
- (2) von Lilienfeld, O. A. First Principles View on Chemical Compound Space: Gaining Rigorous Atomistic Control of Molecular Properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.
- (3) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.
- (4) Balamurugan, D.; Yang, W.; Beratan, D. N. Exploring Chemical Space with Discrete, Gradient, and Hybrid Optimization Methods. *J. Chem. Phys.* **2008**, *129*, 174105.
- (5) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.
- (6) Hoksza, D.; Škoda, P.; Vorsilák, M.; Svozil, D. Molpher: A Software Framework for Systematic Chemical Space Exploration. *J. Cheminf.* **2014**, *6*, 7.
- (7) Hall, R. J.; Mortenson, P. N.; Murray, C. W. Efficient Exploration of Chemical Space by Fragment-Based Screening. *Prog. Biophys. Mol. Biol.* **2014**, *116*, 82–91.

- (8) Weymuth, T.; Reiher, M. Inverse Quantum Chemistry: Concepts and Strategies for Rational Compound Design. *Int. J. Quantum Chem.* **2014**, *114*, 823–837.
- (9) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216.
- (10) Jóhannesson, G. H.; Bligaard, T.; Ruban, A. V.; Skriver, H. L.; Jacobsen, K. W.; Nørskov, J. K. Combined Electronic Structure and Evolutionary Search Approach to Materials Design. *Phys. Rev. Lett.* **2002**, *88*, 255506.
- (11) Tan, D. S. Diversity-Oriented Synthesis: Exploring the Intersections between Chemistry and Biology. *Nat. Chem. Biol.* **2005**, *1*, 74–84.
- (12) Dow, M.; Fisher, M.; James, T.; Marchetti, F.; Nelson, A. Towards the Systematic Exploration of Chemical Space. *Org. Biomol. Chem.* **2012**, *10*, 17–28.
- (13) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvitto, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191–201.
- (14) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120–1127.
- (15) Strasser, S.; Goodman, R.; Sheppard, J.; Butcher, S. A New Discrete Particle Swarm Optimization Algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference, 2016; GECCO '16; ACM: New York, NY, USA, 2016; pp 53–60, DOI: 10.1145/2908812.2908935.*
- (16) Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A. Parallel and Distributed Thompson Sampling for Large-Scale Accelerated Exploration of Chemical Space. 2017, arXiv:1706.01825. arXiv.org ePrint archive. <https://arxiv.org/abs/1706.01825> (accessed Nov 8, 2018).
- (17) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595–10599.
- (18) De Vleeschouwer, F.; Yang, W.; Beratan, D. N.; Geerlings, P.; De Proft, F. Inverse Design of Molecules with Optimal Reactivity Properties: Acidity of 2-Naphthol Derivatives. *Phys. Chem. Chem. Phys.* **2012**, *14*, 16002.
- (19) Pearl, J.; Korf, R. E. Search Techniques. *Annu. Rev. Comput. Sci.* **1987**, *2*, 451–467.
- (20) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613–1623.
- (21) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for de Novo Drug Design - A Survey. *Appl. Soft Comput.* **2015**, *27*, 543–552.
- (22) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134–1145.
- (23) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. PHOENICS: A Universal Deep Bayesian Optimizer. 2018, ArXiv/1801.01469. arXiv.org ePrint archive <https://arxiv.org/abs/1801.01469> (accessed April 4, 2019).
- (24) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495–6502.
- (25) Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Mach. Learn. Res.* **2003**, *3*, 1289–1305.
- (26) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (27) Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.
- (28) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.
- (29) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.
- (30) Teunissen, J. L.; De Proft, F.; De Vleeschouwer, F. Tuning the HOMO-LUMO Energy Gap of Small Diamondoids Using Inverse Molecular Design. *J. Chem. Theory Comput.* **2017**, *13*, 1351–1365.
- (31) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244–13249.
- (32) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; RagHAVACHARI, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09; Gaussian, Inc.: Wallingford, CT, USA, 2009.
- (34) De Vleeschouwer, F.; Chankisijev, A.; Geerlings, P.; De Proft, F. Designing Stable Radicals with Highly Electrophilic or Nucleophilic Character: Thiadiazinyl as a Case Study: Designing Stable Radicals. *Eur. J. Org. Chem.* **2015**, *2015*, 506–513.
- (35) De Vleeschouwer, F.; Geerlings, P.; De Proft, F. Cover Picture: Molecular Property Optimizations with Boundary Conditions through the Best First Search Scheme (ChemPhysChem 10/2016). *ChemPhysChem* **2016**, *17*, 1389–1389.
- (36) Parr, R. G.; Szentpály, L. v.; Liu, S. Electrophilicity Index. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.
- (37) De Vleeschouwer, F.; Speybroeck, V. V.; Waroquier, M.; Geerlings, P.; De Proft, F. An Intrinsic Radical Stability Scale from the Perspective of Bond Dissociation Enthalpies: A Companion to Radical Electrophilicities. *J. Org. Chem.* **2008**, *73*, 9109–9120.
- (38) Hehre, W. J. Ab Initio Molecular Orbital Theory. *Acc. Chem. Res.* **1976**, *9*, 399–406.
- (39) Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822–8824.
- (40) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- (41) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1874.
- (42) Cundari, T. R.; Moody, E. W. Prediction of Bond Dissociation Energies Using Neural Network, Statistical, and Quantum Mechanical Approaches. *J. Mol. Struct.: THEOCHEM* **1998**, *425*, 43–50.
- (43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

- (44) Ramakrishnan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *Chimia* **2015**, *69*, 182–186.
- (45) Hu, X.; Beratan, D. N.; Yang, W. A Gradient-Directed Monte Carlo Approach to Molecular Design. *J. Chem. Phys.* **2008**, *129*, 064102.
- (46) Perone, C. S. Pyevolve: A Python Open-Source Framework for Genetic Algorithms. *SIGEVOlution*. **2009**, *4*, 12–20.
- (47) Head, T.; Loupe, G.; Shcherbatyi, I.; Vinícius, Z.; Schröder, C.; Campos, N.; Young, T.; Cereda, S.; Fan, T.; Shi, K. K. J.; Schwabedal, J.; Santos, C. D. C.; Pak, M.; Callaway, F.; Estève, L.; Besson, L.; Cherti, M.; Pfannschmidt, K.; Linzberger, F.; Cauet, C.; Gut, A.; Mueller, A.; Fabisch, A. *Scikit-Optimize/Scikit-Optimize: v0.5.2*; Zenodo: 2018.