Received: 9 December 2019 | Accepted: 21 April 2020

DOI: 10.1002/cae.22249

RESEARCH ARTICLE

WILEY

Teaching complex molecular simulation algorithms: Using self-evaluation to tailor web-based exercises at an individual level

Oda Dahlen[1] | **Anders Lervik**[1] | **Ola Aarøen**[2] | **Raffaela Cabriolu**[1] | **Reidar Lyng**[3] | **Titus S. van Erp**[1,4]

[1]Department of Chemistry, Faculty of Natural Sciences and Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[2]Department of Biotechnology and Food Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[3]Department of Education and Lifelong Learning, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[4]Center for Molecular Modeling (CMM), Ghent University, Technologiepark, Zwijnaarde, Belgium

Correspondence
Titus S. van Erp, Department of Chemistry, NTNU, Trøndelag, 7941 Trondheim, Norway.
Email: titus.van.erp@ntnu.no

Funding information
Research Council of Norway, Grant/Award Number: 237423; Faculty of Natural Sciences of the Norwegian University of Science and Technology; Olav Thon foundation

Abstract

It is quite challenging to learn complex mathematical algorithms used in molecular simulations, stressing the importance of using the most advantageous teaching methods. Ideally, individuals should learn at their pace and deal with tasks fitting their levels. Web-based exercises make it possible to tailor every small step of the learning process, but this requires continuous monitoring of the learner. Differentiation based on the scores after the first round of common tasks can be demotivating for all students, as they will experience the initial set of tasks as being either too easy or too hard. We designed two tests, a self-monitoring test and a rapid test (RT) in which the students explained equations relating to the current topic. The first test was aimed to see if the students were able to evaluate their own level of knowledge, whereas the RT was aimed to find a fast way to determine the level of the students. We compared both tests with traditional measures of knowledge and used a relatively new method, which was originally designed for the analysis of molecular simulation data, to interpret the results. Based on this analysis, we concluded that self-evaluation, rather than an RT, is a valuable tool to automatically steer individual students through a tree of web-based exercises to match their skill levels and interests.

KEYWORDS

algorithms, rapid tests, self-evaluation, undergraduate education in molecular modeling, web-based exercises

1 | INTRODUCTION

In a university environment, research and teaching are ideally interlinked [19, 20]. Certainly at the master level, it can be highly motivating for both staff and students when parts of the teacher's own contemporary scientific research are transferred to the class. Besides a hopefully contagious enthusiasm that a teacher is likely to transmit to the students, an intertwined approach of teaching and research ensures that state-of-the-art knowledge,

which has not yet arrived in the textbooks, will be provided.

However, when working with complex mathematical methods, it can be challenging to convey your research to others. A consequence can be low student recruitment and an ineffective learning outcome. This is further complicated by the fact that students often have very different backgrounds and skill levels. This stresses the importance of teaching the students in the way they learn best.

The direct motivation for the research performed in this paper is the experience of the university course *Molecular Modeling*, a master course taken by students with a theoretical interest in chemistry, physics, biotechnology, and chemical engineering. In this course, we also aim to report on topics that are a part of our present research focus. In particular, our research group develops algorithms for increasing the accessible time scales of standard molecular simulations. Present time scales of molecular dynamics are generally far from sufficient to study processes like protein folding, chemical reactions, and crystal formation. The algorithms we develop are the so-called rare-event algorithms. More specifically, we develop methods within the *path sampling* family of molecular algorithms that combine two techniques: Monte Carlo and molecular dynamics. Based on a set of clever numerical tricks, depending on the system, the increase in speed, compared with standard molecular dynamics, varies from a modest factor 10 to a few millions. Moreover, a positive aspect of these approaches is that they provide the same results, as an extremely long molecular dynamics simulation, which can be proven using statistical thermodynamics principles [4, 32, 34].

However, both the steps in the algorithmic procedures and the theoretical validation of these steps are quite complex and require special mathematical notation. For this reason, even experienced scientists from the same field have difficulty comprehending the approach by just reading the research articles. The challenge that we face is to make these approaches understandable at the master level using dedicated web applications, ranging from visually instructive videos and interactive web applets to hands-on exercises and questionnaires. It is fairly common in higher education that there is a large variation in prior knowledge and skills in any student cohort, especially in courses that recruit students from several different study programs. To bring all students in a cohort to the point where they can progress from a common ground is a recognized challenge for many teachers. The way to optimally support learning of such complex problems in science, technology, engineering, and mathematics has been the subject of debate and research for many decades [17].

Two major distinctive approaches to this challenge differ most clearly in deciding whether teaching–learning activities for students should be mainly guided or unguided. The first approach is sometimes called non-constructivist, which relies on reducing cognitive load in the design of teaching–learning activities and instructional material [29]. The latter approach belongs to a constructivist understanding of learning including, *that is*, cognitive and social constructivism. This is the basis for a number of teaching–learning designs, such as inquiry-based learning, team-based learning, and problem-based learning [6]. The common aspect of these approaches is that students in some way construct their understanding and learning as a result of interacting with the material at hand, rather than as a result of being told what to learn and understand. There is an ongoing discussion regarding which approach is more successful [13, 14]. Although it seems that novice learners benefit more from direct instructional methods, there is certainly evidence that points out that experienced learners benefit more from a more discovery-based approach [6, 16, 22]. The self-regulated learner [25] is also one of the most important learning outcomes for higher education.

Hence, a pedagogical approach that serves the needs of a diverse group of students should be a mix of guided and unguided teaching–learning activities, and the extent and design of inquiry-based teaching should be tuned to the learning skills of the student. Ideally, each student should be provided with personalized feedback in order to be able to acquire the learning skills necessary for becoming a self-regulated learner. It is well established that feedback that supports the students in constructing an understanding needs to be clear about *what* good performance is, *how* performance so far relates to good performance, and *how* to close the gap [28]. Providing individual feedback is not a viable alternative in many cases. The bigger the cohort, the bigger is the cost in terms of time and resources for providing feedback on an individual basis. This strengthens the need to develop a method of continuously monitoring the level of the students, which does not require more work from the teachers.

Web-based exercises are potentially ideal for providing the learners with a personalized learning experience tailored to their needs during each step of the learning process, but to achieve this, continuous monitoring of the student's level of knowledge is required. One conventional approach is to let all students do the same exercises and give more challenging exercises to the ones who finish quickly. We know that this often leads to boredom among clever students, as they have to go through many easy exercises. Moreover, it can be very demotivating for novice students when they see other students rushing through exercises while they themselves are struggling. Finally, it is also very time consuming, as it requires a

substantial amount of test exercises before a reasonable evaluation of the students' skills can be made.

In this study, we, therefore, investigated alternative ways of monitoring the students with the purpose of differentiation and personalization of learning experiences. With the use of computer-based exercises, this becomes more feasible. We have compared the evaluation of student answers with a series of questions related to molecular simulations in two different ways: first, by checking the answers in a rapid test (RT) covering the same topic, and second, by asking the students to self evaluate the quality of the answer. The RT is commonly used for the purpose described above, whereas the self-evaluation (SE) test is a rather new approach within this context. The use of SE is based on the hypothesis that students can assess their own level and, therefore, tailor the needs of their own learning.

2 | THEORY

Cognitive load theory was proposed by John Sweller in 1988 [30]. Cognitive load refers to the amount of mental effort used in the limited capacity working memory when solving a task. Simply put, extraneous load arises from factors that can be addressed by reducing stress, distractions, or by clarity of information, whereas intrinsic load is imposed by the intrinsic difficulty of the concept that is to be learnt or mastered. The effect of both extrinsic and intrinsic loads may vary considerably from student to student. Later, Sweller et al. [30] further developed the theory to include the concept of germane load, describing the necessary intellectual workload required to process the information at hand for learning. The concept of germane load suggests the possibility of optimizing the design of teaching–learning activities, where good design limits the effects of extrinsic load while taking into consideration the variation in learning progress among students. Several studies in this field outline a balance between direct, or guided, instruction and different kinds of inquiry-based learning [1, 5, 7, 9, 26, 27, 30, 36, 37]. Although it is recognized that children may benefit more in learning from instructional guidance [14, 13], the explicit aim of higher education is to foster students to become self-regulated life-long learners, which implies that students should benefit from being trained in internal guidance as a skill.

In a review of six studies on undergraduates, Clark [6] found that high-ability learners preferred direct instruction, but four out of six studies reported that the students learned more from inquiry-based learning, with no or little difference for the remaining two studies. In contrast, low-ability learners learned more from direct instruction in five of the six studies. Kirschner et al. [13] argued that the students should have acquired a minimum level of knowledge and understanding for self-regulated learning to be superior to receive guided instruction. Also, Roblyer, Edwards, and Havriluk [27] reported that teachers have found that discovery learning is successful only when students have prerequisite knowledge and some prior structured experience.

The challenge, then, is to construct tasks that strike a balance between the needs of novices in the field and those with some experience and prerequisite knowledge, which can support the learning of both the factual knowledge required for higher-order thinking and the development of the same higher-order thinking skills. If the task is too simple, the student will probably do well, regardless of the instruction method, and the conclusion will be that the methods are working equally well. Similarly, we can also expect little difference in the effectiveness of the learning methods if the tasks are too difficult from the start. To tune the level of tasks and the right mix of direct instruction versus inquiry-based learning, some performance feedback is required, ideally without having to expose the students with many ill-suited exercises from the start. Our study compares two approaches to provide feedbacks: first, checking the answers in an RT, and second, asking the students to self-evaluate the quality of their answer.

To analyze the data, we used, among other approaches, the recently developed predictive power method [35] that was originally designed to analyze molecular simulation data, in particular, data from path sampling simulations. The main idea is to check whether certain types of information could be useful to improve the prediction that a certain output function exceeds a predefined threshold, when we already know how often this happens on average. A requirement is that the output function is not the same for all data points, but varying, so that additional a priori information, expressed by certain auxiliary variables, could tell whether the chance to exceed the threshold will be higher or lower than the average. The output function could then be the grade that corresponds to the answers, considering all answers given to all questions and by all students. The output function could also be the average grade for each student, considering all students. The former approach was applied in this article, whereas the RT and the SE results were used as additional information to predict whether the provided answer would exceed a minimum objective score.

3 | METHOD

3.1 | Participants

In total, 18 students participated in the experiment, 10 men and eight women, who were chemistry or

chemical engineers, bachelor or master, students from NTNU. Only one of the students had already followed the *Molecular Modeling* course in which the topics were discussed in detail. However, the type the questions were made was basic enough such that no advance knowledge of molecular simulation techniques was necessary. They could not discuss together and had to sit separately. They were also given maximum two hours to complete the test.

3.2 | Experiment

We constructed a set of 36 questions (see Supporting Information Material) on the basis of some interactive web exercises.[1] The questions were grouped according to the topic. These topics were as follows: *movement of a simple pendulum*, *time steps in molecular simulations*, *movement in a two-dimensional potential*, *reflecting on forces and energies*, *a double-well spring bond*, *calculation of π using random numbers*, and *Replica Exchange Transition Interface Sampling* [33]. These topics consisted of 7, 7, 4, 3, 3, 3, and 9 questions. The last topic was related to our current research, as replica exchange transition interface sampling [RETIS] [33] is a part of the path sampling algorithms, which is being developed in our research group. For each topic (except for *a double-well spring bond*), one of the questions was to write down or explain an equation relating to the topic. This was the RT. For each question, the students were asked to give an answer and then rate their own answer on a scale ranging from 1 to 6. Examples are shown in Figure 1.

In the exercises related to Figure 1a, we asked questions related to Newton's equation of motion and how a timestep affects a simulation, specifically regarding the error. We gave equations $\vec{x}(t+\Delta t) = \vec{x}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)\Delta t^2$ and $\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{\vec{a}(t)+\vec{a}(t+\Delta t)}{2}\Delta t$, and asked them to explain Δt, $\vec{x}(t+\Delta t)$, $\vec{v}$ and $\vec{v}(t+\Delta t)$. In the middle picture, we asked how the forces of the particles change, what is the relationship between the forces and the velocities or acceleration of the particles, and what it is called when a particle reaches the edge of the dashed square, disappears, and reappears on the opposite side. In the last picture, we asked if they could explain what the picture depicted and also what the line drawn in the image meant in the context of a chemical reaction.

[1]The web applets were derived from the PyRETIS [15] program, which is freely available at pyretis.org.

3.3 | Measures

The objective evaluation (OE), was obtained by correcting the exercises without looking at the students' self-rating. The average grade of the OE of the students was 3.75, which corresponds to a C according to a European credit transfer system grading scale. The SE was the students' own judgment of their answer, which was measured on the same scale (1–6). The RT consisted of six questions, each question belonging to one of the topics. The result of the RT was compared with the average OE (excluding the RT question) for the specific topic. Some of the questions (see Supporting Information Material) were considered to require knowledge or skills not reflected by the RT. These questions were, therefore, not considered in the analysis in which the outcome was compared with the RT evaluations.

3.4 | The Pearson product–moment correlation coefficient

The Pearson correlation coefficient [24] is a number describing the strength of a linear association between two variables. It is measured by drawing a linear plot with one of the variables along the x-axis and the other one along the y-axis and, then, finding the best fit through the points. This coefficient, called r, then indicates how far are all these points from the best linear fit. The Pearson correlation coefficient can range from +1 to −1, and a value of zero means that there is no linear association between the variables. A value above zero implies a positive linear correlation, whereas a value below zero implies a negative linear association. The equation for r can be expressed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (1)$$

where x_i and y_i are the i-th value of x and y in the samples, respectively. $\bar{x}$ and $\bar{y}$ are the means or averages of the samples.

3.5 | Predictive power method

Another way to analyze the results, which we applied to the test results, was actually inspired by the subject of the test itself. As stated before, the main aim of this study is to improve the teaching of complex molecular simulation algorithms, especially path sampling methods. Recently, we developed an analysis method, called the predictive

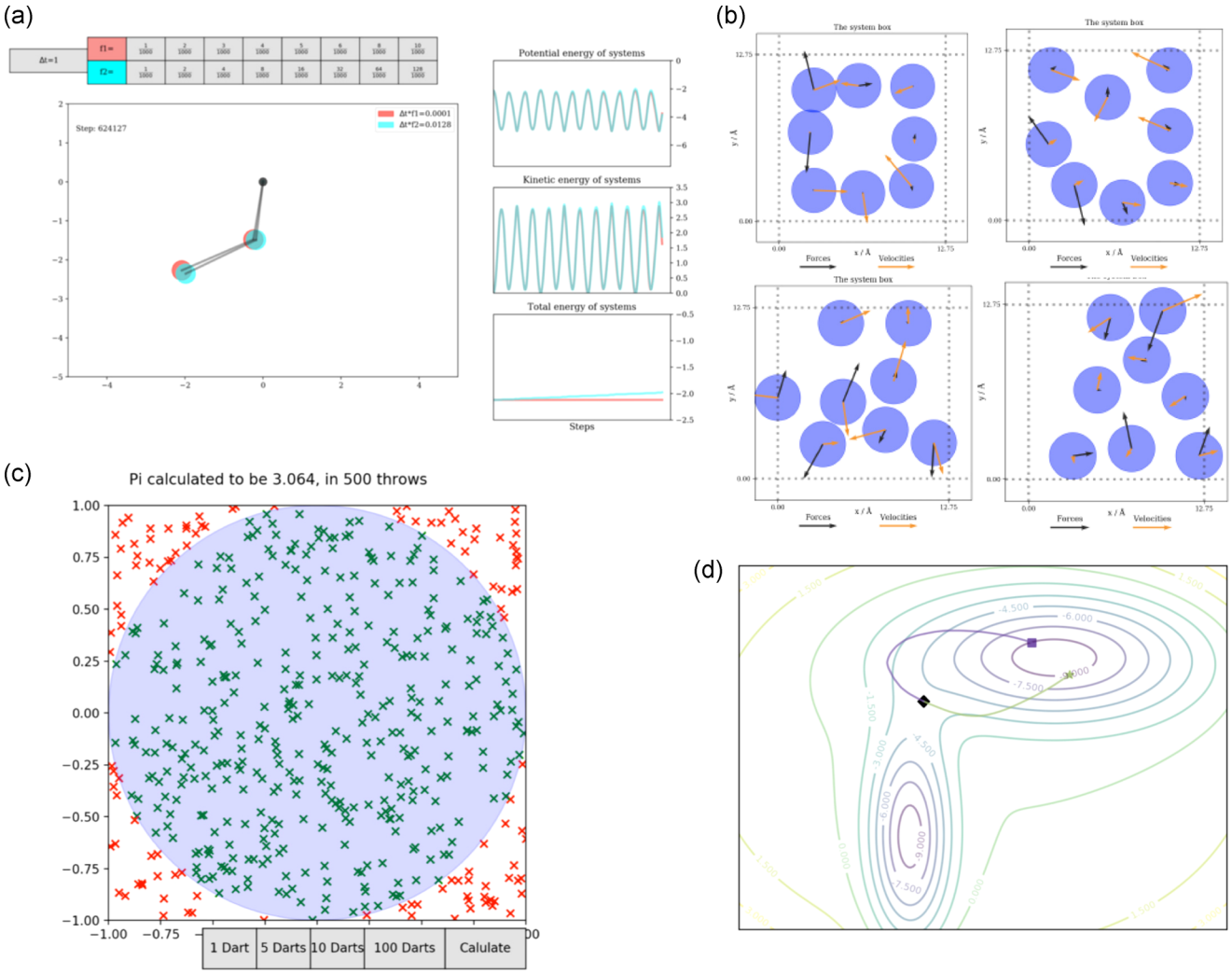

FIGURE 1 Examples of four exercises given in the experiment. Panel a shows a screenshot of the exercise on *time steps in molecular simulations*, b shows a screenshot of the exercise on *movement in a two-dimensional potential, reflecting on forces and energies*, c shows a screenshot of the exercise on *Calculate pi using random numbers*, and d shows a screenshot of the exercise on replica exchange transition interface sampling

power method [35], to detect hidden variables in the data of path sampling simulations that correlate to reactivity. The strength of the correlation is expressed by a measure called the *predictive capacity*.

If we exploit the analogy between *reactivity* and *obtaining a minimal score* on the test or on a single question, the exact same machinery of analysis can be used. In the molecular version of the predictive power method, molecular trajectories are categorized into *reactive* trajectories or *unreactive trajectories*, based on a progress coordinate. Trajectories reaching a minimal threshold progress are assigned as reactive and the others as unreactive. The predictive power method aims to identify the auxiliary conditions that are not already described by the progress coordinate, which can predict whether a trajectory is reactive or not at an early stage, that is, when the progress coordinate is still below the threshold [18]. Similarly, we can try to identify which parameters correlate with getting a correct or wrong answer.

We here give a label "r" (reactive) or "u" (unreactive) to each student answer if OE, respectively, exceeds or not exceeds a score OE^r. One can also decide to consider only those answers that reached a minimum score OE^c (here, "c" stands for crossing [35]) to remove nonserious attempts or questions that were unanswered due to time trouble. Naturally, OE^c must be chosen to be smaller than OE^r. Hence, each answer is assigned as "r" if $OE \geq OE^r$ and as "u" if $OE^c \leq OE < OE^r$. Let q be another variable or a set of variables different from OE. Here, q can be the SE (the SE given by the student), RT (the RT connected to the specific question), or

combinations. The q parameter can even be nonnumeric, like q identifying which student delivered the answer. An example of how this is computed is given in the Supporting Information Material, Section 2. Naturally, the SE and RT are natural ways of measurement to be used in a web-based exercise.

The predictive power method for this application works as follows. An auxiliary variable q should be chosen. All questions should be considered, and the values for q and score OE should be calculated. Then, for a chosen threshold OE^r and minimum score OE^c, distributions $r(\cdot)$ and $u(\cdot)$ should be constructed:

$$r(q'|\mathrm{OE}^r, \mathrm{OE}^c) = \frac{\text{number of answers with } q = q' \text{ and OE} \geq \mathrm{OE}^r}{\text{total number of answers with OE} \geq \mathrm{OE}^c} \text{ and } u(q'|\mathrm{OE}^r, \mathrm{OE}^c) = \frac{\text{number of answers with } q = q' \text{ and } \mathrm{OE}^c \leq \mathrm{OE} < \mathrm{OE}^r}{\text{total number of answers with OE} \geq \mathrm{OE}^c}.$$

Now, let $P(\mathrm{OE}^r, \mathrm{OE}^c)$ be the total fraction of all answers with a score larger than or equal to OE^c, with $\mathrm{OE} \geq \mathrm{OE}^r$.

$$P(\mathrm{OE}^r, \mathrm{OE}^c) = \frac{\text{number of answers with OE} \geq \mathrm{OE}^r}{\text{total number of answers with OE} \geq \mathrm{OE}^c}.$$

In path simulations, this term is called the *crossing probability* [31].

The following relations must hold:

$$\begin{aligned} &\sum_q [r(q|\mathrm{OE}^r, \mathrm{OE}^c) + u(q|\mathrm{OE}^r, \mathrm{OE}^c)] = 1, \\ &\sum_q r(q|\mathrm{OE}^r, \mathrm{OE}^c) = P(\mathrm{OE}^r, \mathrm{OE}^c), \\ &\sum_q u(q|\mathrm{OE}^r, \mathrm{OE}^c) = 1 - P(\mathrm{OE}^r, \mathrm{OE}^c), \end{aligned} \tag{2}$$

where the summations must be taken over all possible q values.

If q is a perfect indicator of score OE being larger than OE^r, we would expect that $u(q'|\mathrm{OE}^r, \mathrm{OE}^c) = 0$, for any value q', whenever $r(q'|\mathrm{OE}^r, \mathrm{OE}^c) > 0$ and vice versa, $r(q'|\mathrm{OE}^r, \mathrm{OE}^c) = 0$ whenever $u(q'|\mathrm{OE}^r, \mathrm{OE}^c) > 0$. This means that if the value of q is known, we will know with 100% certainty whether the score exceeds OE^r or not. In that case, the predictive capacity $T(\mathrm{OE}^r, \mathrm{OE}^c)$ should be equal to 1. The other extreme case is that whenever q has absolutely no informative value about the OE. This means that our predictions with the additional information on q are just as good as for a random question of a random student; the chance to exceed OE^r, given a minimum score of OE^c, is simply $P(\mathrm{OE}^r, \mathrm{OE}^c)$, and the predictive capacity should have the same value. Most practical cases will lie in between these two extremes, for which we define the predictive capacity as [35]

$$T(\mathrm{OE}^r, \mathrm{OE}^c) = \sum_q \left\{ \frac{r(q|\mathrm{OE}^r, \mathrm{OE}^c)}{(r(q|\mathrm{OE}^r, \mathrm{OE}^c) + u(q|\mathrm{OE}^r, \mathrm{OE}^c))} \times \left[\frac{r(q|\mathrm{OE}^r, \mathrm{OE}^c)}{\sum_{q'} r(q'|\mathrm{OE}^r, \mathrm{OE}^c)} \right] \right\}. \tag{3}$$

It should be note that $r(q|\mathrm{OE}^r, \mathrm{OE}^c)/[r(q|\mathrm{OE}^r, \mathrm{OE}^c) + u(q|\mathrm{OE}^r, \mathrm{OE}^c)]$ is the probability for having $\mathrm{OE} > \mathrm{OE}^r$ for a specific value of q. The predictive capacity reflects a weighted average of these probabilities for different q values. The weight is given within the squared brackets and can be viewed as the relevance of the q value for the r distribution; it is large for those q values that describe a large part of the $\mathrm{OE} \geq \mathrm{OE}^r$ answers and small otherwise. Equation (3) can be rewritten using Equation (2) as [35]

$$T(\mathrm{OE}^r, \mathrm{OE}^c) = 1 - \frac{1}{P(\mathrm{OE}^r, \mathrm{OE}^c)} \times \sum_q \frac{r(q|\mathrm{OE}^r, \mathrm{OE}^c) u(q|\mathrm{OE}^r, \mathrm{OE}^c)}{r(q|\mathrm{OE}^r, \mathrm{OE}^c) + u(q|\mathrm{OE}^r, \mathrm{OE}^c)}, \tag{4}$$

where the last part after the minus sign can be viewed as the overlap between the $r(\cdot)$ and $u(\cdot)$ distributions.

If the overlap is small, it implies that the two distributions in q space are well separated and q is apparently a good parameter to discriminate between the data points belonging to the "reactive" or the "unreactive" category (i.e., for our case, answers having an objective score $\geq \mathrm{OE}^r$ and having a score $< \mathrm{OE}^r$). Hence, if one does not know the value of the OE of an answer, but knows the value of the corresponding q variable and knows beforehand how the distributions $u(q)$ and $r(q)$ appear, it is possible to predict very well whether $\mathrm{OE} \geq \mathrm{OE}^r$ or not. This quality of prediction is reflected by the small overlap, which implies that the T-function is nearly 1. If, however, the ratio between $r(q)$ and $u(q)$ is the same along the full q-axis, even if the absolute values of $r(q)$ and $u(q)$ may vary, then the overlap will be $1 - P(\mathrm{OE}^r|\mathrm{OE}^c)$ [35], and thus $T(\mathrm{OE}^r|\mathrm{OE}^c) = P(\mathrm{OE}^r|\mathrm{OE}^c)$. More explanation about how to compute the predictive ability $T(\mathrm{OE}^r|\mathrm{OE}^c)$ is given in the Supporting Information Material.

4 | RESULTS AND DISCUSSION

In total, the students answered 594 out of 648 questions. This means that each student, on average, answered 33

out of 36 questions, and each question, on average, was answered by 16.5 out of the 18 students. There was no significant difference between men and women. Figure 2 shows the difference between the SE and OE for each student and each question.

It is clear that the results nearly cover the full scale ranging from −5 to 5. The extreme values occur in the case of a student estimating an answer to be absolutely wrong, when in fact it is fully correct (maximal underestimation), or analogously for the reverse situation of maximal overestimation. Both extremes occurred equally often, 11 and 10 times. Also, 41% of the data points show exact correspondence (SE–OE = 0). The difference between SE and OE per student and per question is shown in Figure 3.

The sample mean per student was −0.13 and the standard deviation was 0.76 (here we mean the actual standard deviation of the answers, not the standard deviation of the mean that relates to the statistical error or 68% confidence interval). The difference SE–OE per question had a sample mean of −0.16 and an *SD* of 0.59. From this, we can observe that the students, on average, evaluated themselves very close to the OE, despite a considerable number of large overestimations and underestimations. As both occur equally often, they get canceled in the average.

Figure 4 shows the relationship between the OE and SE (top), and OE and the RT (bottom). The height of the column at (x, y) is a count of the number of times OE = x and SE = y at the same time. In the top image, we observe from the column heights that OE = SE = 1, OE = SE = 6, and OE = 5, SE = 6 are occurring 30, 15, and 20 times as often as the average of the other columns, respectively.

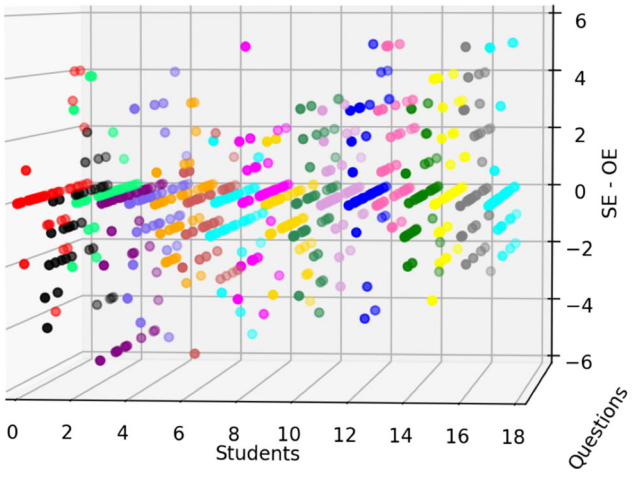

FIGURE 2 Plot showing the self evaluation–objective evaluation (SE–OE) for each student and each question. The total number of data points in this plot is 594, of which 243 are at the surface, SE–OE = 0, implying that the question was evaluated correctly

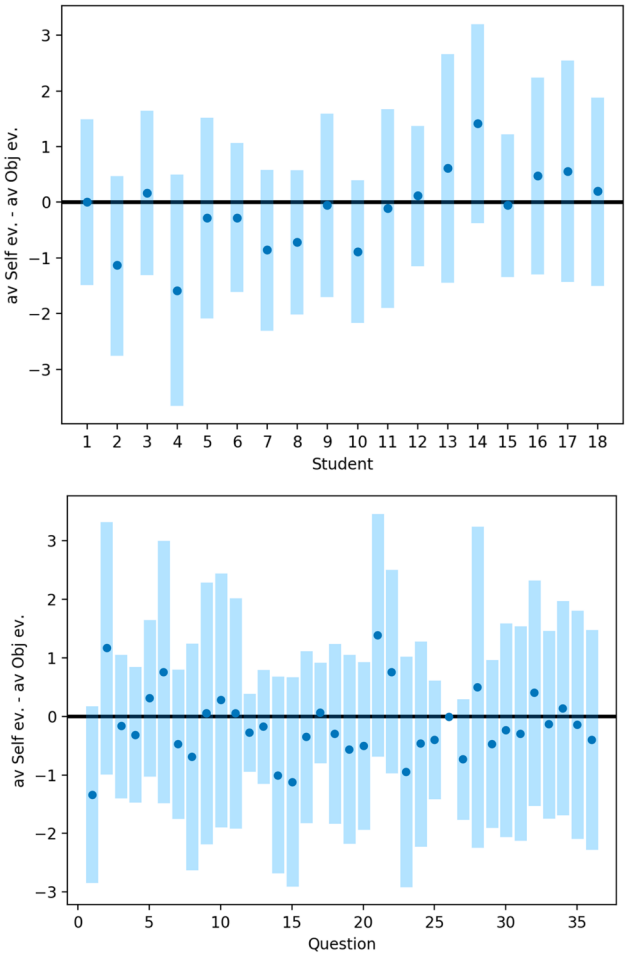

FIGURE 3 (Top) Plot showing the average deviation, SE–OE, for each student, 1–18. (Bottom) average deviation, SE–OE, for each question, 1–36. The height of the bars (from center to top/bottom) indicates the standard deviation of the data for which the average is computed. OE, objective evaluation; SE, self evaluation

Evaluating yourself when you are sure that you either have the correct answer or the wrong answer seems to be much easier than evaluating yourself when you are on a grade between 2 and 5. In the bottom image, we observe that that the histograms in the OE versus RT are more spread as compared with the top picture. Specifically, the extremes SE = 1, RT = 6, and SE = 6, RT = 1 have higher heights than other columns, as compared with the top panel.

The Pearson product–moment correlation coefficient for the OE versus SE was calculated from Equation (1), and r was found to be 0.627 indicating that the relation is to some extent linear. For the relation of the OE as a function of the RT, we found the Pearson product–moment correlation to be $r = 0.543$. Hence, also this shows approximate linear correlation, though somewhat lower than for the OE versus SE dependence. Indeed, Figure 4, shows that there is a larger fraction of counts in off-diagonal bars in the OE versus RT graph compared to OE versus SE. Though the

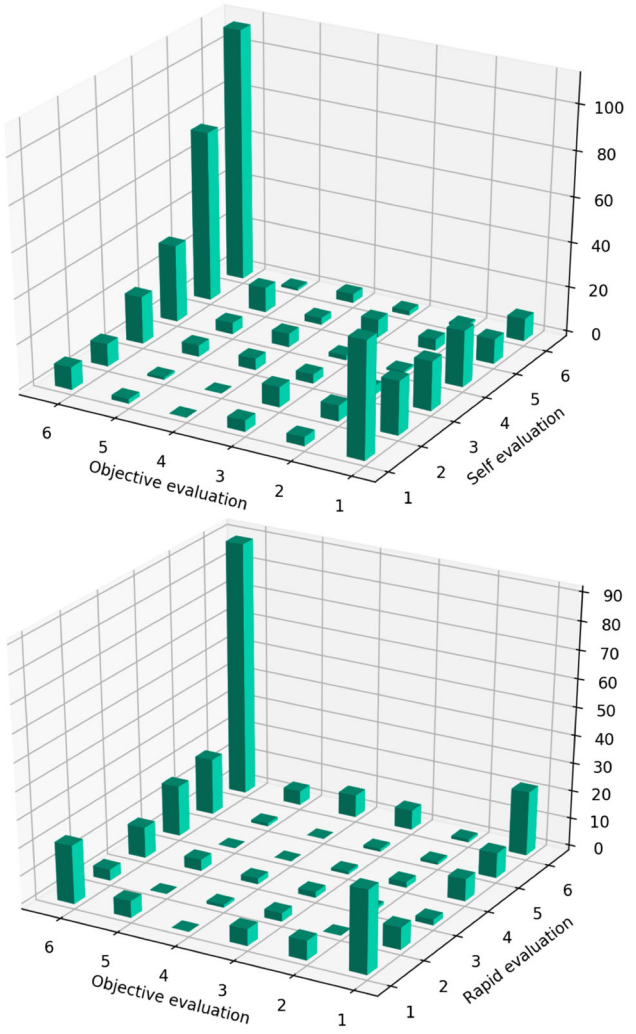

FIGURE 4 3D plot of the correlation between the OE and SE (top) and OE and RT (bottom). In the calculation of the correlation between OE and SE, the questions from the rapid test was excluded from the sample. OE, objective evaluation; SE, self evaluation

linear character in both graphs is predominantly due to the high corner values at OE = SE equal to 1 and 6. Hence, it is rather difficult to make conclusions based on the difference in the Pearson product–moment correlation coefficients as both curves are not linear enough to make a quantitative assessment.

Finally, we applied the predictive power analysis [35] using as auxiliary parameters $q = \text{SE}$, $q = \text{RT}$, $q = (\text{OE}, \text{RT})$ and $q = \text{Stud}$ (student). The latter refers to the case where the auxiliary information is nonnumeric but refers to the student who provided the answer (see Supporting Information Material). The use of multidimensional parameter (like $q = (\text{OE}, \text{RT})$) or nonnumeric parameter (like $q = \text{Stud}$) is evidently not possible to study with the Pearson method. The minimum score OE^c was set equal to 1 in all of our analysis. This implies that all answers were considered except when nothing was written down (unanswered questions, mainly due to time limitations as reported by the students). Also, the RT questions were not considered for the averages of the OE score. Instead, RT was only used as an auxiliary variable; for each answer with a score ≥ 1 and not part of the RT questions, we determined the OE, and the auxiliary parameters $q = \text{SE}$ and $q = \text{RT}$. Here, the SE value was the SE by the student of that specific answer and the RT value was the score in the RT that corresponded to the non-RT question for which the answer was given. Hence, all answers by a student within a specific theme, for which an RT was designed, were assigned to the same RT value. The results of $P(\text{OE}^r, \text{OE}^c = 1)$ as a function of OE^r is given in Figure 5a together with $T(\text{OE}^r, \text{OE}^c = 1)$ for $q = \text{SE}$, RT, (SE, RT), and Stud.

Clearly, $T(\text{OE}^r, \text{OE}^c)$ is always equal or strictly larger than $P(\text{OE}^r, \text{OE}^c)$ but the increase is significantly larger for $q = \text{SE}$ than for $q = \text{RT}$. The combined analysis $q = (\text{SE}, \text{RT})$ further enhances the predictive power, but only slightly. The difference in predictive performance of the different q variables becomes more evident if we visualize the relative difference of $T(\text{OE}^r, \text{OE}^c)$ compared to $P(\text{OE}^r, \text{OE}^c)$ as is done in Figure 5b. The SE score is a considerably better predictor for the OE than the knowledge of who actually gave the answer ($q = \text{Stud}$) as is illustrated by their relative predictive capacities which is twice as large for $q = \text{SE}$ than for $q = \text{Stud}$. This immediately implies that SE is a better predictor than any person inherent specific variable like IQ of the student, hours of preparation by the student (not relevant for this test), background knowledge of the student, etc. The RT variable, on the contrary, show a relative predictive capacity that is twice as low as the one for $q = \text{Stud}$.

The difference in predictive capacity between $q = \text{SE}$ and $q = \text{RT}$ is further analyzed in Figure 6, where the $u(\cdot)$ and $r(\cdot)$ distributions are shown for the two cases, SE and RT, for $\text{OE}^r = 4$, 5, and 6. The distributions for the different OE^r values look very similar. The distributions $u(\text{SE})$ and $r(SE)$ both have a single maximum, but at opposite extremes. The distributions $u(\text{RT})$ and $r(RT)$, on the contrary, are double peaked. We will look further into the $\text{OE}^r = 6$ case.

As we see in Figure 6a, $r(\text{SE}|\text{OE} = 6)$ is nearly zero at SE = 1, whereas $u(\text{SE}|\text{OE} = 6)$ peaks here. More quantitatively, these values are 0.02 and 0.126 which implies that if the student gave a SE equal to 1, it is nearly certain $(0.126/[0.126 + 0.02] = 86\%)$ that the OE will be less than 6. Reversely from the values of $u(\cdot)$ and $r(\cdot)$ at SE = 6 (0.036 and 0.222, respectively), we see that the chance of a score OE equal to 6, given the SE score was 6, is also 86% $(0.222/[0.222 + 0.036])$. However, given the higher absolute values $(r(\text{SE=6}) + u(\text{SE=6}) = 0.222 + 0.036 = 0.258 > r(\text{SE=1}) + u(\text{SE=1}) = 0.126 + 0.02 = 0.146)$, it

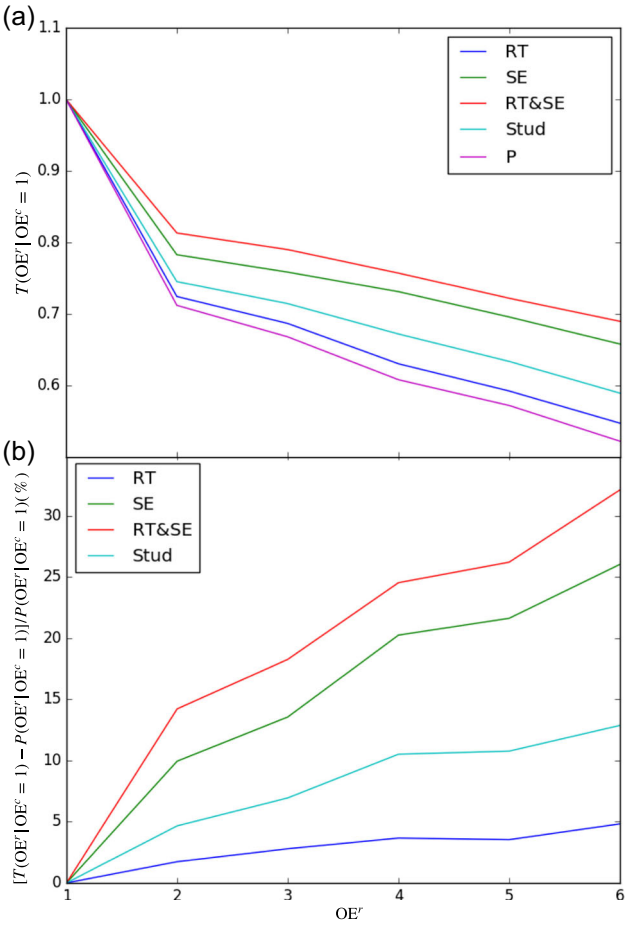

FIGURE 5 (a) Predictive capacity $T(\mathrm{OE}^r,\ \mathrm{OE}^c)$ with fixed $\mathrm{OE}^c = 1$ as a function of OE^r is shown for auxiliary variables q = SE, RT, (SE, OE), and Stud. As reference also $P(\mathrm{OE}^r, \mathrm{OE}^c)$ is shown in the same plot. The minimum value OE^c implies that everything except unanswered questions were included in the analysis. (b) Same data relative to $P(\mathrm{OE}^r, \mathrm{OE}^c)$. OE, objective evaluation; SE, self evaluation; RT, rapid test

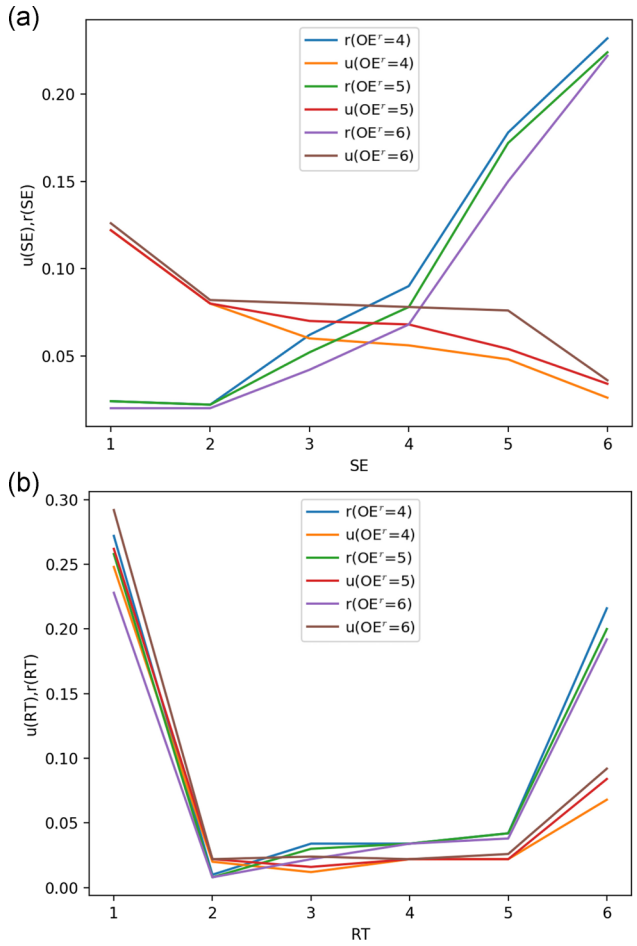

FIGURE 6 Distributions of $u(q)$ and $r(q)$ are shown for (a) q = SE and (b) q = RT. SE, self evaluation; RT, rapid test

is clear that it happens more often that students evaluate their answers with the maximal score than with the minimal score.

The double-peaked character of the $u(\mathrm{RT})$ and $r(\mathrm{RT})$ distributions is due to the type of RT questions whose answers are mostly either fully correct or completely wrong with little possibility to score something in between a 1 and a 6. Again, the distributions look similar for the different OE^r values considered (corresponding to the distributions $\mathrm{OE} \geq 4$, $\mathrm{OE} \geq 5$, and $\mathrm{OE} = 6$). Again we focus on $u(\cdot)$ and $r(\cdot)$ for the $\mathrm{OE}^r = 6$ case and examine the values at RT = 1 and 6. For RT = 1, $u(1) = 0.292$ which is only slightly higher than $r(1)$ that is 0.228. Hence, scoring a very bad RT is not very predictive with respect to having an OE equal to 6 or having an OE lower than that. These chances are nearly equal. Note that also $P(6, 1) = 0.522$ implying that someone scoring RT = 1 has still nearly the same chance to obtain the maximal OE score, as someone scoring an RT > 1. Reversely for RT = 6, we have $u(6) = 0.092$ and $r(6) = 0.192$ implying that with a maximal score in the RT test the chance to get also a maximal OE becomes twice as large than to get a lower OE. Still, from the absolute numbers $r(1) = 0.228$ and $r(6) = 0.192$, we can conclude that among the answers with a maximum OE score, there were more answers having a corresponding RT = 1 than a RT = 6 evaluation. The predictive capacity for obtaining the maximum OE, $T(\mathrm{OE}^r = 6, \mathrm{OE}^c = 1)$ can be expressed as a weighted average of the $r(q)/[r(q) + u(q)]$ probabilities for all possible q values $(q = 1, ..., 6)$ where the weights are proportional to $r(q)$. So, even if the fact that RT = 6 is a reasonable predictor for OE = 6, $(0.192/[0.192 + 0.092] = 68\%)$, it is given a lower weight than the RT = 1 case in the computation for overall the predictive capacity. The predictive power method establishes hence a measure that reflects a measure of predictivity of individual q values that is balanced with the number of occurrences of these q values within the r-ensemble.

We, therefore, believe that we convincingly demonstrated that the SE is a better predictor than the RT and that the predictive power method provides a more conclusive analysis than the Pearson product–moment correlation method with respect to this quantitative comparison between RT and SE (Figure 5b). Naturally, one has to be careful as our analysis is based on a limited number of volunteers and the test results are dependent on the phrasings of both the OE and the RT questions.

Although there are aspects of our particular design of questions in this study, the findings that SE has a better effect on the students' learning than an RT is in agreement with a wide body of work. Nicol and Macfarlane-Dick [21] presented seven principles for self-regulated learning and feedback that support and develop self-regulation on students. Based on previous work [3, 8, 10, 11, 12], they discussed several problematic aspects of formative feedback: first, that feedback needs to be more than mere information about whether the student is right or wrong, if the student's self-regulation skills are to be developed; second, that the feedback is easily decoded by the student; and, third, that considering feedback only to be about information neglects that feedback interacts with the student's motivation and beliefs. Finally, if feedback is seen only as correcting information for the student, then the workload of providing that information grows with the number of students. Using RTs that can be corrected automatically addresses the last of these issues but falls short of addressing the first three. The first two principles of good feedback practice, among the seven identified by Nicol and Macfarlane-Dick [21], are helping to clarify what good performance is, and facilitating the development of self-assessment. Both these are to some degree satisfied at least by the combination of OE and SE. The RT cannot in its present design convey more information than whether the student is right or wrong, akin to checking for the correct answer. In future work we shall further address the design of both the SE and the RT with the ambition to support both the students' learning of the necessary factual knowledge and the development of self-regulating learning skills.

5 | CONCLUSION AND PROSPECTIVES

The results presented in this paper belong to a first experimental stage of our new teaching approach to enhance students learning. A fundamental challenge for students is to build the bridge between the mathematical description and the real chemical-physics process. It has been established that the integration of information-communication-technology tools in traditional teaching methods can greatly enhance student learning in higher education [23]. In this context, we propose web-based exercises integrating SE methods and statistical analysis aiming on quantitatively measuring progresses and faults in the learning process of the students. Web-based exercises are able to create a personalized learning experience that fits the needs and time of the individual student. The self-evaluating tool encourages each students to think on her/his own performance. In addition, it allows the students to detect their weak points and strategies for improvement. Finally, the SE together with automated procedure to correct numerical exercises could be used to make a selection from a large database of exercise to offer the best training.

To quantify the value of SE in teaching molecular simulations via web-based exercises, we developed a small test consisting of 36 basic questions and asked 18 volunteering students to solve them. After completion of the evaluation, we examined the relations among the scores of the SE, the RT, and the final OE. Analysis based on the Pearson product–moment correlation coefficient was not conclusive due a lack linearity. We, therefore, applied the relatively new analysis method that originated from the molecular simulation field itself. The predictive power method [35] is an analysis method derived for path sampling simulations to identify auxiliary parameters that can be used to predict whether a reaction will take place or not. In this case, we examined whether knowing the RT and SE scores are predictive with respect to obtaining an OE larger than a threshold OE^{r}. The advantage of the predictive power method is that it does not require a near-linear correspondence between the two variables being compared (RT and SE) and the OE. The resulting analysis showed, more convincingly than the Pearson product–moment correlation analysis, that the SE is a better predictive parameter than RT for obtaining a high OE score.

A plausible explanation for this finding is that the SE and OE values reflect a measure of exactly the same question, whereas RT reflects a measure of a different, though related question. On the contrary, RT and OE have in common that both reflect evaluations done in an objective manner by the teacher, whereas SE is not. Our predictive power analysis shows that the first connection is stronger than the second one. Naturally, these findings are very much dependent on the question design. But it can be very challenging if not impossible to design a question that is fast but still faithfully tests the skill and knowledge levels required for a more elaborate question.

In the coming years, we plan to develop a database of questions for the *Molecular Modeling* course and test if we can effectively use the SE tool to automatically select at an individual level the best matching tasks. Given the nature

of molecular simulations as a scientific field, which already involves extended use of numerics and computer visualizations, training using web-based exercises is a natural development. Moreover, we plan to intertwine the series of web exercises with links to instruction videos and animations such that theory of a certain subject could be refreshed before commencing with a new set of questions. We believe that this approach will be applicable to many other courses which build up on a strong numerical or mathematical foundation and could well be combined with the "flipped-classroom" pedagogical approach [2].

ACKNOWLEDGMENTS

The authors thank the Olav Thon foundation for supporting them with the development of teaching within an academic environment and the students that volunteered in the experiment. In addition, the authors are grateful to the Research Council of Norway for financial support for the development of simulation software and research (project number 237423) and the Faculty of Natural Sciences of the Norwegian University of Science and Technology to support development in educational methods.

ORCID

Oda Dahlen http://orcid.org/0000-0002-1048-0605
Anders Lervik http://orcid.org/0000-0002-1505-6731
Ola Aarøen http://orcid.org/0000-0002-1317-2042
Raffaela Cabriolu http://orcid.org/0000-0002-9859-7444
Titus S. van Erp http://orcid.org/0000-0001-6600-6657

REFERENCES

1. M. Albanese and S. Mitchell, *Problem-based learning: A review of literature on its outcomes and implementation issues*, Acad. Med. **68** (1993), 52–81.
2. J. Bergmann, and A. Sams, Flip your classroom: Reach every student in every class every day, International Society for Technology in Education, Washington DC, 2012.
3. D. Boud, *Sustainable assessment: Rethinking assessment for the learning society*, Stud. Contin. Educ. **22** (2000), no. 2, 151–167.
4. R. Cabriolu et al., *Foundations and latest advances in replica exchange transition interface sampling*, J. Chem. Phys. **147** (2017), no. 15, 152722.
5. Z. Chen and D. Klahr, *All other things being equal: Acquisition and transfer of the control of variables strategy*, Child Dev. **70** (1999), 1098–1200.
6. R. E. Clark, *Antagonism between achievement and enjoyment in ati studies*, Educ. Psychol. **17** (1982), no. 2, 92–101.
7. J. Colliver, *Effectiveness of problem-based learning curricula: Research and theory*, Acad. Med. **75** (2000), 259–66.
8. C. Dweck, Self-Theories: Their Role in Motivation, Personality and Development, Psychology Press, Philadelphia PA, 1999.
9. R. S. Eisenstaedt, W. E. Barry, and K. Glanz, *Problem-based learning: Cognitive retention and cohort traits of randomly selected participants and decliners*, Acad. Med. **65** (1990), S11–S12.
10. T. Garcia, *The role of motivational strategies in self-regulated learning*, Understanding self-regulated learning (P. R. Pintrich, ed.), Jossey-Bass, San Francisco CA, 1995.
11. R. Higgins, P. Hartley, and A. Skelton, *Getting the message across: The problem of communicating assessment feedback*, Teach. High. Educ. **6** (2001), no. 2, 269–274.
12. R. Ivanic, R. Clark, and R. Rimmershaw, *What am I supposed to make of this? The messages conveyed to students by tutors' written comments*, Student writing in higher education: new contexts (M. R. L. B. Stierer, ed.), Open University Press, Buckingham UK, 2000.
13. P. A. Kirschner, J. Sweller, and R. E. Clark, *Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching*, Educ. Psychol. **41** (2006), no. 2, 75–86.
14. D. Klahr and M. Nigam, *The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning*, Psychol. Sci. **15** (2004), no. 10, 661–667.
15. A. Lervik, E. Riccardi, and T. S. van Erp, *PyRETIS: A well-done, medium-sized python library for rare events*, J. Comput. Chem. **38** (2017), no. 28, 2439–2451.
16. T. Loveless, H. F. Ladd, and C. Rouse, *The use and misuse of research in educational reform*, Brook. Pap. Educ. Policy **1** (1998), 279–317.
17. National Research Council. R. A. McCray, R. L. DeHaan, and J. A. Schuck (eds.), *Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop*, The National Academies Press, Washington, DC, 2003.
18. M. Moqadam et al., *Local initiation conditions for water autoionization*, Proc. Natl. Acad. Sci. USA **115** (2018), E4569–E4576.
19. R. Neumann, *Perceptions of the teaching-research nexus: A framework for analysis*, High. Educ. **23** (1992), no. 2, 159–171.
20. R. Neumann, *The teaching-research nexus: Applying a framework to university students' learning experiences*, Eur. J. Educ. **29** (1994), no. 3, 323–338.
21. D. Nicol and D. Macfarlane-Dick, *Formative assessment and self-regulated learning: A model and seven principles of good feedback practice*, Stud. Higher Educ. **31** (2006), no. 2, 199–218.
22. G. R. Norman and H. G. Schmidt, *Effectiveness of problem-based learning curricula: Theory, practice and paper darts*, Med. Educ. **34** (2000), no. 9, 721–728.
23. R. Oliver, *The role of ICT in higher education for the 21st century: ICT as a change agent for education* (2003). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.9509
24. K. Pearson, *Note on regression and inheritance in the case of two parents*, Proc. Royal Soc. London **58** (1895), 240–242.
25. P. R. Pintrich and A. Zusho, *Student motivation and self-regulated learning in the college classroom*, *Higher Education: handbook of theory and research* (vol. XVII) (J. C. Smart and W. Tierney, eds.), Agathon Press, New York NY, 2002.
26. E. Pollock, P. Chandler, and J. Sweller, *Assimilating complex information*, Learn. Instruction **12** (2002), no. 1, 61–86.
27. M. Roblyer, J. Edwards, and M. Havriluk, *Integrating educational technology into teaching*, The University of Texas, Prentice Hall, Boston, MA, 1997.

28. D. R. Sadler, *Formative assessment and the design of instructional systems*, Instructional Sci. **18** (1989), no. 2, 119–144.
29. J. Sweller, *Cognitive load during problem solving: Effects on learning*, Cogn. Sci. **12** (1988), no. 2, 257–285.
30. J. Sweller, V. J. J. G. Merrienboer, and F. G. W. C. Paas, *Cognitive architecture and instructional design*, Educ. Psychol. Rev. **10** (1998), no. 3, 251–296.
31. T. S. van Erp, D. Moroni, and P. G. Bolhuis, *A novel path sampling method for the sampling of rate constants*, J. Chem. Phys. **118** (2003), 7762–7774.
32. T. S. van Erp and P. G. Bolhuis, *Elaborating transition interface sampling methods*, J. Comput. Phys. **205** (2005), 157–181.
33. T. S. van Erp, *Reaction rate calculation by parallel path swapping*, Phys. Rev. Lett. **98** (2007), no. 26, 268301.
34. T. S. van Erp, *Dynamical rare event simulation techniques for equilibrium and nonequilibrium systems*, Adv. Chem. Phys. **151** (2012), 27.
35. T. S. van Erp et al., *Analyzing complex reaction mechanisms using path sampling*, J. Chem. Theory Comput. **12** (2016), 5398.
36. D. Vernon, L. Blake, and R. Does, *Problem-based learning work? a meta-analysis of evaluative research*, Acad. Med. **68** (1993), 550–63.
37. B. H. Verhoeven et al., *An analysis of progress test results of PBL and non-PBL students*, Med. Teach. **20** (1998), 310–316.

AUTHOR BIOGRAPHIES

Oda Dahlen obtained a Master of Technology (M. Tech.) in nanotechnology at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. She started a PhD position in 2014 where she studied rare event simulation techniques, dynamics of mesoscopic DNA models, in combination with didactical aspects of teaching complex molecular algorithms at the master level. She defended her Ph. D. Thesis in 2019 and is presently working as information technology consultant.

Anders Lervik is an Associate Professor in the Department of Chemistry, Norwegian University of Science and Technology (NTNU). He received his M.Tech./M.Sc. in Physical Chemistry in 2003 from the Department of Chemistry (NTNU) in 2008, and a Ph.D. degree in Chemistry in 2012 from the Department of Chemistry, NTNU. His research interests include non-equilibrium thermodynamics, molecular simulations, rare-event simulations with applications to biological systems, interfacial transport, transport in porous media, phase transformations, and catalysis.

Ola Aarøen is a PhD candidate from the NTNU where he is studying coalescence with optical tweezers at the Department of Biotechnology and Food Science. Before that, he received a Master's degree in Applied Theoretical Chemistry in 2016 at the same university and was hired at a research internship for the development of interactive web applications that were used in the NTNU master course Molecular Modeling.

Raffaela Cabriolu is a scientist at the Laboratory of Molecular Simulation of EPFL, Switzerland. She has been awarded a Ph.D. in Physics in 2012 from the School of Physics and Astronomy at the University of Leeds, UK. She received her Master's Degree in Physics in 2009 by the Physics Department of the University of Cagliari, Italy. She is a computational physicist with a broad interest in statistical physics and in the modeling of chemical-physics processes. In particular, she has used computer simulations to investigate porous materials, network forming materials in glassy and crystalline phases, phase transitions, supra-molecular aggregation in biological systems, and, in general processes taking place as a rare (activated) process, such as nucleation events.

Reidar Lyng is Associate Professor of university pedagogics at the Department of Education and Lifelong Learning, at NTNU, presently chairing the Center for Science & Engineering Education at NTNU, Trondheim, Norway. He is a MSc in Chemical Engineering and holds a PhD degree in Physical Chemistry. He has more than 30 years of experience of education development from NTNU and several Swedish universities. His research and development interests are wide ranging and include the systemic interplay between teachers, students, and learning spaces.

Titus van Erp received a Master's Degree in physics from the Radboud university in Nijmegen, the Netherlands, in 1999 and his Ph. D. from the University of Amsterdam, the Netherlands, in 2003. After postdocs at the ENS-Lyon, France, and the University of Leuven, Belgium, he became Associate Professor at the NTNU in 2012,

where he got promoted to Full Professor in 2016. His primary research activities are the application and development of advanced simulation algorithms within physics, chemistry, and biology.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Dahlen O, Lervik A, Aarøen O, Cabriolu R, Lyng R, van Erp TS. Teaching complex molecular simulation algorithms: Using self-evaluation to tailor web-based exercises at an individual level. *Comput Appl Eng Educ*. 2020;28:779–791. https://doi.org/10.1002/cae.22249