

RSC Advances



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. This *Accepted Manuscript* will be replaced by the edited, formatted and paginated article as soon as this is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.



Cite this: DOI: 10.1039/xxxxxxxxxx

TD-DFT based fine-tuning of molecular excitation energies using evolutionary algorithms

Sailesh Abburu,^a Vishwesh Venkatraman,^a and Bjørn K. Alsberg^{a*}

Received Date

Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

An evolutionary *de novo* design method is presented to fine-tune the excitation energies of molecules calculated using time-dependent density functional theory (TD-DFT). The approach is applied to a π -conjugated molecular system, azobenzene. The excitation energies for all the molecules generated by the evolutionary design scheme were computed at TD-DFT level on multiple supercomputing clusters. A software developed in-house was used to automatically set up the TD-DFT calculations and exploit the advantages of parallelization and thereby speed up the process of obtaining results for the evolutionary *de novo* program. Our proposed optimisation scheme is able to propose new azobenzene structures with significant decrease in excitation energies.

1 Introduction

Molecular and materials design using computers has been in practice for many years and has evolved over time. The increased availability of high performance computing resources has facilitated the design of novel materials without the need to perform experiments¹. However, the task of discovering new molecules that satisfy multiple criteria still remains a challenge. In recent years many research groups have performed quantum chemical calculations on massively networked computing systems to accelerate the discovery of new materials. For instance, Hachmann *et al.*² have developed a virtual high-throughput screening algorithm to create a large database of molecules for potential use as materials in photovoltaic cells. In the project, extensive quantum chemical calculations for more than 2 million structures (combinatorially generated from a set of building blocks) were performed on a grid computing network to evaluate properties related to the power conversion efficiency. Using a similar strategy, de Jong *et al.*³ implemented a scheme to calculate the elastic properties of inorganic crystalline compounds using density functional theory (DFT) methods and developed a database of such molecules. In a recent article, Korth⁴ investigated the performances of semi-empirical, DFT and wave function based methods for applications to screen a large number of molecular structures relevant for battery electrolytes. In a related study, Qu *et al.*⁵ computed the ionisation potentials and electron affinities

of about five thousand molecules using DFT methods, relevant for battery electrolyte systems. These applications of DFT based high-throughput virtual screening (HTVS) methods were primarily aimed at development of a database of molecules for access to the scientific community.

The search for structures with desired properties is often based on various global optimisation algorithms such as genetic algorithms^{6–11} (GA) and particle swarm optimisation¹² (PSO) techniques. These methods have been used extensively in computer-aided drug design^{13–17} and protein ligand-docking^{18–20} studies. Both of these optimisation routines are metaheuristic and require many iterations to reach acceptable solutions. The central theme in these optimisation routines is to generate new structures from a pool of molecules discovered in earlier iterations. Since evolutionary routines (such as GAs) rely on the property estimates, i.e. fitness of known/discovered molecules, it is essential to estimate the fitness as accurately as possible. While property estimates using experimental methods would be the most accurate, it is extremely time consuming and expensive⁶. In computer-aided molecular design (CAMD), the fitnesses are generally estimated using *knowledge-based*^{21–23} scoring functions or *empirical*^{24–26}. In knowledge-based scoring functions, commonly employed in protein-ligand docking studies, the interaction energies are computed as inverse formulation of the Boltzmann law. The technique is used to identify atom-pairs that are preferred while penalising atom-pairs that exhibit repulsive interaction. Empirical scoring functions are obtained by using supervised regression techniques where, experimentally observed property measures are approximated as a function of certain chemical or structural features of the molecule. Although these methods are fast and compu-

^a Department of Chemistry, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway. E-mail: alsberg@ntnu.no

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

tationally cheap, they do not always provide a reasonably accurate estimate of the true fitness²⁷. To address this, we present a more reliable DFT based fitness evaluation algorithm that is fully automated and scalable, designed to leverage on the processing power of multiple computers on a distributed network. Owing to this, the algorithm allows for fitness calculations of many molecules simultaneously. All aspects of the fitness evaluation, starting from generation of the input files, parsing of the output data and launching required computations are automatically managed by our software.

We apply this high-throughput (HT) fitness evaluation scheme in the evolutionary *de novo* design of molecules with lower excitation energies. The lowest energy of excitation (here defined as the λ_{max}), is an important property in many fields of applications, such as organic electronics and photovoltaics. For instance, in dye sensitised solar cells (DSSCs), research has shown that dyes with a λ_{max} closer to the infra-red region tend to have higher power efficiency^{28–34}. In areas like organic light emitting diodes (OLEDs) and sensor applications, it is oftentimes desirable to have the HOMO-LUMO energy levels within a narrow range in order for the molecule to be suitable candidates for applications^{35–37}. A design scheme that allows for fine control of the these energy levels is therefore desirable.

As an application example we have chosen to apply our *de novo* based design scheme to azobenzenes. We selected azobenzenes because they are small and often are structurally symmetric, making them tractable for DFT and TD-DFT computations. Additionally, numerous applications of azobenzenes exploiting its photoisomerisation behaviour have been reported in the scientific literature^{38–46}. Azobenzenes fall in an interesting class of molecules, that exhibit a change in their geometric shape (*cis-trans* isomerization) when irradiated with light of appropriate wavelength. This allows us to control certain physical properties of the molecule using light, and hence, making it possible to control processes *in vivo*⁴⁷. Depending on the field of application, the photoisomerisation property of azobenzenes is exploited using light at appropriate wavelength. The traditional trial and error approach to discover structures with the desired property, has mostly been time consuming and tedious^{38–41}. Thus, a more rational design approach is needed to improve selected properties. In an earlier attempt to identify such azobenzenes, Carstensen *et al.*²⁶ used GAs for design of azobenzenes that could photoisomerise using commonly available lasers. The UV-Vis absorbance spectra were computed using semi-empirical methods. We expand on their work, by implementing a more accurate DFT and TD-DFT based computation of the UV-Vis spectra of the molecules. The goal in our evolutionary *de novo* design is to discover azobenzenes that have λ_{max} closer to the infra-red region of the electromagnetic spectra, i.e low excitation energies. Although we have chosen to maximise the λ_{max} (minimise the excitation energy), the method could also have been used to fine-tune the λ_{max} to a target value.

In evolutionary design, the starting point of the optimisation is generally a set of randomly generated molecules, called the initial population. The algorithm improves the fitness (a figure of merit for each molecule) of the molecules starting from this

set of structures. A randomly generated initial pool may not have fitness measures better than pre-existing molecules and is also unlikely to include those molecules. Hence, starting the design scheme from a random initial pool makes it a suboptimal strategy for optimisation. In our *de novo* design scheme we, therefore, use an evolutionary algorithm (EA) that allows for the inclusion existing and known molecules in the initial population, such that improvement in the fitness starts from existing molecules with high fitness values.

2 Methods

2.1 Fragment based design

In chemistry, molecules can be viewed as a set of substructures (fragments) linked to each other in a defined manner. One can view the substituent groups attached to the common substructure (scaffold) as the cause for the different physical property values. For example, the red or blue shift of the UV-Vis absorbance peaks in azobenzenes (see Figure 1), based on the substituent groups attached to the scaffold. In ethanol, unsubstituted azobenzene (Mol 1 in Figure 1) has its λ_{max} at 318 nm, while 4-[(E)-2-(4-methanesulfonylphenyl)diazen-1-yl]-N,N-dimethylaniline (Mol2 in Figure 1), has its λ_{max} at 445 nm. This means properties of azobenzene (and its derivatives) can be tuned by modifying the substituents bonded to the scaffold. This fragment based design of molecules is the most common approach in computational materials and molecule design⁴⁸. Our *de novo* evolutionary program follows this design routine to discover molecules with improved λ_{max} measures.

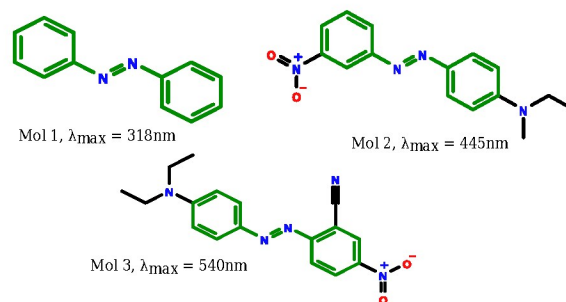


Fig. 1 Azobenzene derivatives, with a common scaffold (in green) and substituent groups attached leading to a change in the observed λ_{max} .

2.2 Evolutionary *de novo* design

The *de novo* design process flow is shown in Figure 2. The design scheme begins by generating a set of random structures or by including a set of pre-existing molecules (defined by the user) in the initial population. New molecules are built by attaching fragments to scaffolds that are randomly selected from a predefined library. In the next step the fitnesses of all molecules in the initial population are computed. The fitness evaluation step is generally not required when an existing set of structures is included in the initial pool that already have associated fitnesses. Following this, all molecules in the population are ranked based on their fitness measures. The algorithm then selects a subset of the structures from the population to which evolutionary opera-

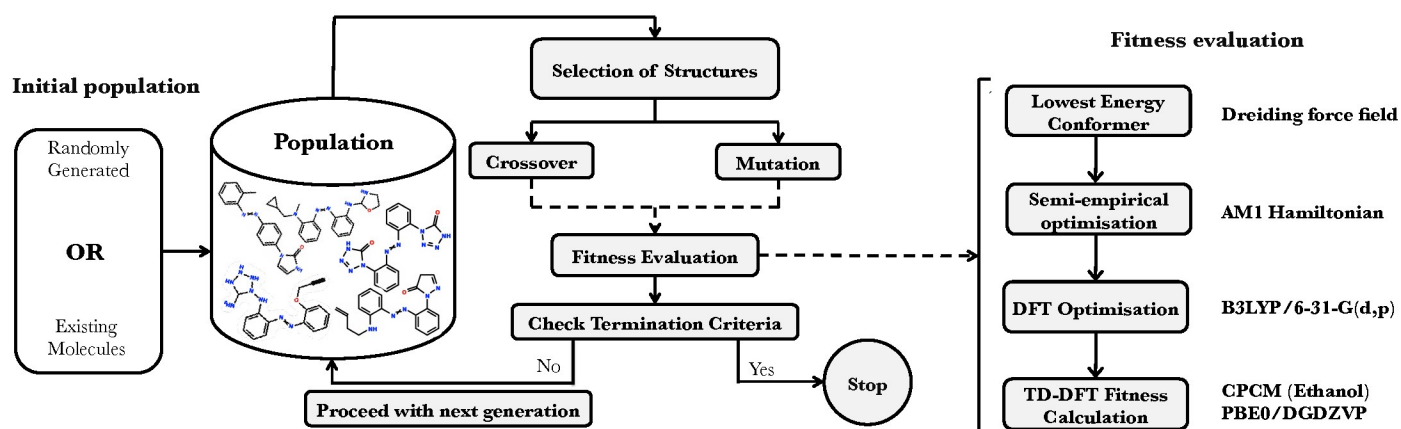


Fig. 2 Process flow of our evolutionary *de novo* design method. Steps in fitness evaluation is shown in the right.

tors such as crossover and/or mutation are performed. In contrast to standard genetic algorithms (GAs)^{6,49} which uses a linear genetic material, our evolutionary *de novo* design approach represents the molecules as graphs which are very similar to the classical 2D graph representation of molecules in chemistry. In this sense, the *de novo* method used here more resembles genetic programming^{7,11,50} (GP) than classical GAs.

Figure 3 shows the evolutionary operators used in our *de novo* design. In the crossover operation, fragments between two molecules (usually, the best two molecules) are swapped, whereas in the mutation operation, one or more fragment group(s) in a molecule are either added, deleted or substituted by a randomly selected fragment from the library. These changes to the structures are brought about by taking synthetic compatibility rules into account (discussed in section 2.3). These genetic operations mimic natural evolutionary process and their use is the key step in the design process. The process of modifying selected structures using genetic operators and obtaining new molecules with improved property, is probabilistic. The new molecules are then subject to the fitness evaluation, following which, all molecules in the population are re-ranked based on their fitness measures. This process of selecting structures and performing genetic operations is repeated until the number of cycles defined (no. of generations) is completed.

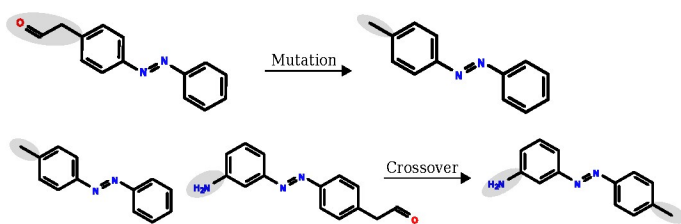


Fig. 3 Genetic operators schemes, mutation and crossover. The substituent groups highlighted in grey are the fragments involved in the respective genetic operation.

2.3 Synthetic accessibility

To maximise the probability of the structures from the *de novo* design being synthesizable, specific rules to link fragment and scaffolds were established. These rules control which atoms in the fragment group and scaffold can be linked. In addition, it also specifies the associated reaction class to bring about the chemical modification. To generate synthetically realistic structures a suitable fragment library was created. For all fragments in the library, information about the attachment atom(s) on the substructure, the bond order and the corresponding reaction class were defined. The reaction class depends on the chemical environment of the attachment atom. For example, all fragments having α -hydrogens can theoretically undergo aldol condensation⁵¹ reaction with an OH containing fragment, forming a C-C bond at the α -carbon atom. Hence α -carbon atoms in all fragments can have the same reaction class (aldol condensation). All reaction classes and their mutual compatibility rules are specified in a table, which is read by the evolutionary *de novo* software to generate synthetically realistic structures.

The quality of the molecular fragments provided by the user is equally important for successful search of realistic structures with improved properties. One way to improve the quality of the fragments is by fragmenting specific bonds in existing molecules, to give synthetically realistic substructures. In this work, the fragment library was customised for the molecular system being studied. To achieve this, specific bonds in about 300 existing azobenzenes were broken to obtain fragments and scaffolds. Figure 4 shows the fragmentation routine followed to generate the different fragments. The molecules were fragmented using a Python program developed in-house. Fragmenting existing molecules in this fashion, allowed the generation of scaffolds with attachment points that follow similar reactions. More precisely, attachment atoms at *ortho* and *para* positions in the benzene rings generally get assigned the same reaction class, which falls in line with the theory of aromatic substitution reactions⁵¹. The substructures obtained were then added to the respective fragment and scaffold libraries after a check for duplicates was done.

However, having fragments only from existing structures may

confine the search space. Hence, to ensure sufficient diversity, the fragment library was augmented with additional 700 structures. These 700 structures were selected by applying a molecular weight constraint (maximum of 150 g/mol) on a larger pool with 23,000 structures obtained from the BRICS⁵² (Breaking of Retrosynthetically Interesting Chemical Substructures) fragment database. Fragmenting molecules in this fashion also allowed for inclusion of all three hundred azobenzenes in the initial population.

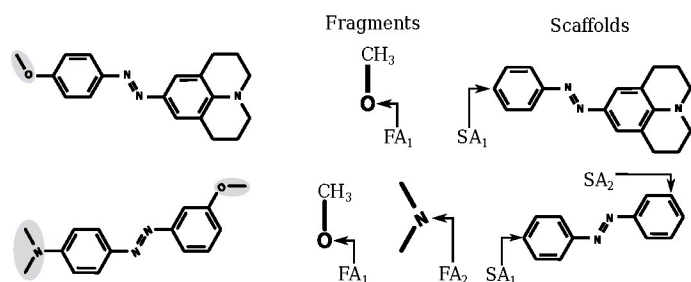


Fig. 4 Fragmentation of azobenzene molecule to give fragments and scaffolds. Each group of atoms highlighted (in grey) form a fragment. The rest of the atoms form the scaffold, retaining the same attachment atom points. FA_i and SA_i , ($i = 1, 2$) are fragment and scaffold attachment atoms.

2.4 Fitness evaluation

As shown in figure 2, the fitness evaluation was performed in four steps. To compute the λ_{max} , the geometry of a molecule was first optimised and then its UV-Vis absorbance spectrum was computed. The first three steps relate to the geometry optimisation of the molecule at different levels of theory. In the first step, a search for the lowest energy conformer was done using molecular mechanics. This was followed by semi-empirical and DFT level geometry optimisations (in gas phase) of the molecule in the second and third step respectively. In the final step, the optimised coordinates of the molecule was used to compute its absorbance spectrum using TD-DFT calculations in the solvent phase.

The lowest energy conformer search of the molecule was done using *cxcalc*⁵³ (from the ChemAxon software package), based on the Dreiding force field⁵⁴. The geometry of the lowest energy conformer was then optimised using the Austin Model 1⁵⁵ (AM1) Hamiltonian, implemented in the MOPAC⁵⁶ software package. This was followed by DFT level optimisation of the geometry in gas phase using the NWChem software⁵⁷. The density functional and basis set used in the DFT optimisation was Beckes three-parameter and Lee-Yang-Parr hybrid (B3LYP)⁵⁸ and 6-31G(d,p)⁵⁹ respectively. Although, solvent phase geometry optimisation offers the best chances of minimising errors, due to computational difficulties associated with solvent-solute interactions, the optimisations were performed in gas phase. The optimised geometry in gas phase was considered a sufficiently close approximation to the solvated structure⁶⁰. TD-DFT spectra computations were performed in ethanol using *Gaussian 09*⁶¹ program. Solvation effects were included by using the conductor-like polarisable continuum model (CPCM)⁶². The density functional and basis set used for TD-DFT calculations were the PBE0⁶³ and double- ζ -valence plus

polarisation (DGDZVP)⁶⁴ respectively.

2.5 Software framework for fitness evaluation

The *de novo* software was designed to accommodate independent fitness evaluation routines in the optimisation. Here, the first two steps in the fitness evaluation were performed on the desktop computer where the *de novo* program was run. Computationally demanding DFT and TD-DFT calculations were done on two supercomputing clusters. To automate this process, a dedicated high-throughput (HT) fitness evaluation algorithm was developed and written in Python. The program allows DFT and TD-DFT based fitness assessment of many molecules simultaneously. The HT fitness program was designed to perform all the practical tasks related to the process such as, to and fro transfer of files to supercomputing clusters and execution of DFT and TD-DFT computations. The input file for various steps in the fitness evaluation scheme were prepared using Open Babel⁶⁵ application which was called through the Python program. The progress of these computations on all the computing resources was monitored by the Python script. All the required files from the remote computers were retrieved once the jobs were completed. Figure 5 outlines

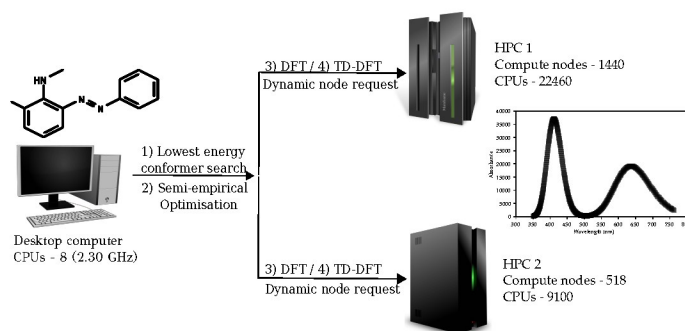


Fig. 5 Allocation of DFT/TD-DFT jobs to supercomputers based on the number of compute nodes available on them.

the algorithm followed to automate the process of fitness evaluation. The lowest energy conformer and semi-empirical optimisation steps were performed on a desktop computer. The remaining steps in the fitness evaluation were performed on supercomputers. The HT fitness program was designed to make optimal use of the computing resources available. The decision to perform calculations on a supercomputing cluster was made based on the number of free compute nodes available. In addition, the number of nodes required for the tasks was decided on the fly. At this point, it should be noted that the supercomputing clusters are a shared resource and therefore, number of nodes available on them changed. Once the jobs on the supercomputers were launched, the Python routine checked for the status of a job every 3 minutes. The outputs of the DFT and TD-DFT program were then copied back to the desktop computer. This system enabled centralised control over all the computing resources used in this study. If any of the steps in the fitness evaluation failed, the calculations for that molecule were terminated and a new molecule was considered.

2.6 Computational details

The evolutionary *de novo* design software has been developed in-house and is written in Java using the CDK⁶⁶ toolkit. The crossover and mutation probabilities were set to 0.65 and 0.35 respectively. An additional constraint on the molecular weight (max. 550g/mol) of molecules was added to restrict the size of the molecules, thereby increasing the probability of the molecule being synthesizable. The population size was set to 100 and the number of offsprings to be produced in each iteration was set to four. The maximum number of iterations (generations) was set to 100. In the TD-DFT spectra, for a transition to be considered significant, the minimum oscillator strength cut-off was set to 0.1⁶⁷. The *de novo* program was run on a desktop computer with eight intel i7-3615QM CPU cores and each CPU core having maximum clock speed of 2.30 GHz. The system specifications of the two supercomputer clusters HPC 1 and HPC 2 are shown in Figure 5. HPC1 is an SGI Altix ICE X distributed memory system and has 1440 compute nodes⁶⁸. All nodes in the cluster have two 8-core Intel Sandy Bridge (2.6 Ghz) cores and 32 GB memory. HPC2 is an HP BL 460c Gen 8 cluster having a total of 518 compute nodes and 9132 Intel E5-2670 cpu cores each with 32 GB memory⁶⁹.

3 Results and discussion

3.1 TD-DFT exchange functional and basis set

The absorbance spectrum of a molecule can be computed using several quantum chemical methods like TD-DFT, complete active space methods (CAS-PT2)⁷⁰ and equation-of-motion coupled-cluster scheme (EOM-CC)⁷¹. While CAS-PT2 and EOM-CC are often more accurate than TD-DFT, their computational expense is considerable and hence their applicability is limited^{60,72} in HT screening. Challenges associated with these methods in the calculation of the absorbance spectra are further amplified when solvent phase computations are to be performed⁷³. TD-DFT, on the other hand, manages to produce comparable results at a much lower computational cost even for solvent phase computations⁷⁴. However, a large variance can be seen in the TD-DFT calculated spectra when using different density functional and basis-set combinations. The accuracy of the predicted spectra using a certain density functional and basis-set combination, varies with the molecular system being studied^{60,75,76}.

To determine the best density functional for the molecular system in our study, the λ_{max} was computed using three widely used density functionals⁷⁶. These estimates were then compared with experimental data. The basis set, DGDZVP⁶⁴ was used to compute the UV-Vis absorbance spectra for all three functionals. The three DFT functionals investigated were PBE0⁶³, B3LYP⁵⁸ and CAM-B3LYP⁷⁷. The DFT functional with lowest mean absolute error (MAE) in the λ_{max} estimates was selected for the TD-DFT absorbance spectra computation of molecules from the *de novo* design. About 70 molecules (test) were chosen from a larger set of around 300 azobenzenes with experimental λ_{max} reported⁷⁸. The experimental λ_{max} measures of the test set ranged from 322 nm to 575 nm (see ESI). The fitness evaluation routine described in section 2.4 was followed to predict the spectra of the test set molecules. In the final step, i.e., TD-DFT spectra computa-

tion was performed using three different functionals for the same molecule. As shown in Figure 6, the MAE in λ_{max} predicted by PBE0⁶³ functional was 18 nm, while the same for B3LYP⁵⁸ and CAM-B3LYP⁷⁷ functionals were 26 nm and 44 nm respectively. Furthermore, the error in prediction using PBE0 functional was nearly consistent in all three groups of molecules as shown in Table 1. Based on this study, PBE0 density functional was chosen as the best functional to calculate the TD-DFT spectra for the azobenzenes in our study.

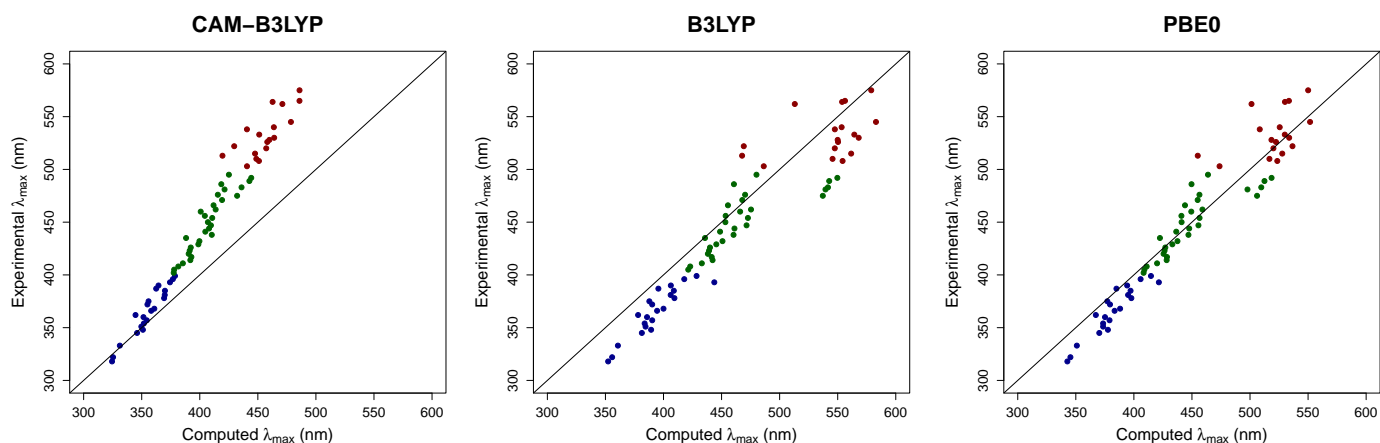
3.2 Azobenzenes with improved fitness

In an evolutionary design approach, the initial population forms the basis for the subsequent generations. The selection of molecules for evolutionary operators depends on the ranking of the molecules in the initial population. In our *de novo* design implementation the initial population can either be randomly generated by the GA or can be provided by the user. The latter method is generally preferable when there exists a set of already discovered molecules and their fitness property measure is known. In this study, both approaches were implemented while all other parameters related to the genetic algorithm, such as the crossover and mutation probability, the scaffold and fragment library and restriction on the maximum molecular weight of the molecules, were kept constant. To compare the performances of both the methods, the maximum of the fitnesses in the initial population, i.e. 575 nm, was used as the benchmark to assess the quality of the initial population.

In the first run, a set of about 300 known azobenzene molecules with their experimentally measured fitness (λ_{max}) was supplied to the GA program as the initial population. The fitness of this set of molecules ranged from 318 nm to 575 nm. The first molecule with a λ_{max} above the benchmark (575 nm), was discovered in the first generation. This was followed by two more such molecules in the sixth generation and fourth and the final molecule with a fitness measure higher than the benchmark, was discovered in generation no. 53 with a λ_{max} of 608 nm. The four molecules (MOL1-MOL4) discovered using this GA set up are shown in Table 2.

In the second run, the initial population of molecules were randomly generated by the GA code. The highest λ_{max} in the initial population, was computed to be 505 nm and the lowest λ_{max} computed was 332 nm. As expected, the results with this *de novo* setting were quite different from the earlier setting. The first molecule with a λ_{max} absorbance peak above 575 nm was obtained in generation no. 44. The next molecule obtained with a λ_{max} above 575 nm was in generation no. 83, with a λ_{max} of 580 nm. The two molecules (MOL5-MOL6) discovered using this method are shown in Table 2.

Table 3 shows the transitions of all the six molecules at their respective absorbance peaks. Transitions of five molecules (except MOL4) at their absorbance peaks is dominated by HOMO→LUMO transitions (close to 100%), which shows that these peaks correspond to their lowest excitation energies. From these results it can be seen that, the first molecule with a λ_{max} higher than 575 nm (the benchmark set) was discovered in the first generation in



(a) $R^2 = 0.95$, MAE = 42 nm, $\sigma_{Pred.Error} = 30$ nm (b) $R^2 = 0.86$, MAE = 26 nm, $\sigma_{Pred.Error} = 16$ nm (c) $R^2 = 0.92$, MAE = 18 nm, $\sigma_{Pred.Error} = 18$ nm

Fig. 6 Plots of Experimental λ_{max} vs Computed λ_{max} of seventy one azobenzenes using different functionals (6a, 6b, 6c). Squared correlation between experimental λ_{max} and computed λ_{max} (R^2), mean absolute error in prediction (MAE) and standard deviation in prediction error ($\sigma_{Pred.Error}$) for all functionals are mentioned in the plots. The blue, green and red points indicate molecules with experimental λ_{max} ranging from 300-400 nm, 400-500 nm and 500-600 nm respectively.

Table 1 Table below showing the variation in correlation coefficient between experimental λ_{max} and computed λ_{max} (R^2), mean absolute error in prediction (MAE) and standard deviation in prediction error ($\sigma_{Pred.Error}$) with different class of molecules. Based on the experimental λ_{max} measures, 71 molecules were grouped into three classes, exp. λ_{max} range 300-400 nm (22 molecules), 400-500 nm (31 molecules) and 500-600 nm (18 molecules).

Exp. λ_{max} range	300-400 nm (22 Mols.)			400-500 nm (31 Mols.)			500-600 nm (18 Mols.)		
	R^2	MAE (nm)	$\sigma_{Pred.Error}$	R^2	MAE (nm)	$\sigma_{Pred.Error}$	R^2	MAE (nm)	$\sigma_{Pred.Error}$
CAM-B3LYP	0.96	11.0	8.0	0.93	41.5	13.5	0.76	77	14.2
B3LYP	0.91	27.7	10	0.80	22	18	0.40	29	16
PBE0	0.92	16.3	8	0.85	13	10	0.46	20	18

the first run, while in the second run it was identified in the 44th generation. It was observed that in the first run, 4 DFT and TD-DFT computations were required to discover the first molecule with an improved property. In the second run it took a significantly larger number of DFT and TD-DFT calculations (460) to identify the first molecule with desired property (i.e. $\lambda_{max} > 575$ nm). Also the number of molecules with improved measure of property discovered in the first and second *de novo* runs were, four and two respectively.

3.3 Discussion

In an evolutionary design scheme like ours, many parameters influence the outcome of the results. For example, the crossover and mutation probabilities, in the GA set up can to some extent control the diversity in population. In this study these parameters were arbitrarily assigned. A high crossover probability can lead to structures that are not very different from the existing molecules, whereas a high mutation probability increases the diversity of the population. Optimal probabilities for these genetic operators can be determined by many experiments. The results in both the *de novo* design approaches, a plateauing effect of the fitness was seen, from which the system seem unable to escape. Possible reasons for this could be, non-optimal genetic operator probabilities, restricted fragment diversity and inaccuracies in the fitness estimates.

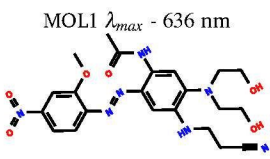
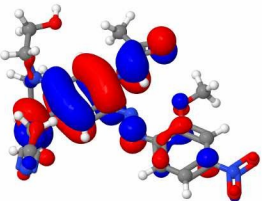
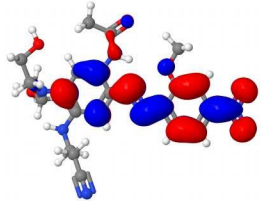
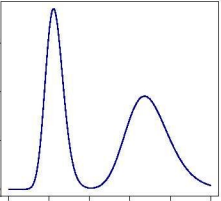

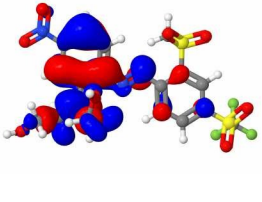
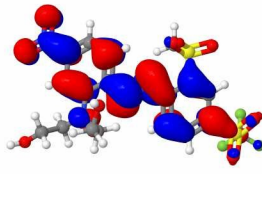
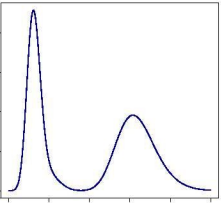
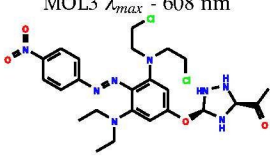
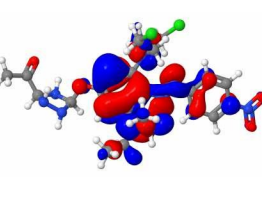
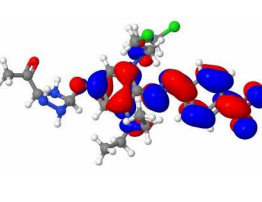
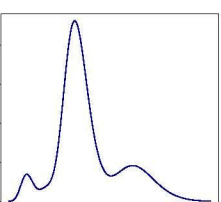
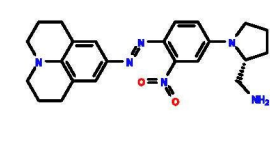
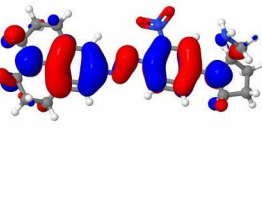
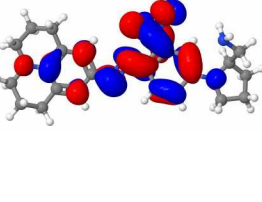
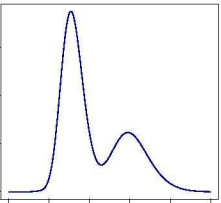
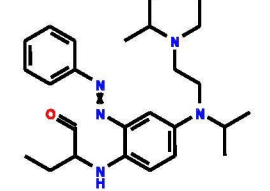
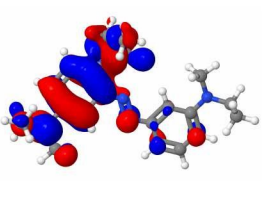
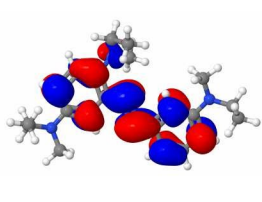
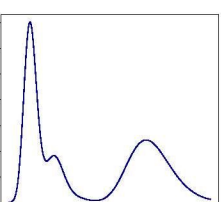
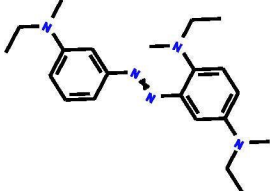
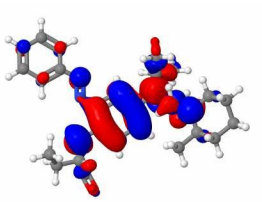
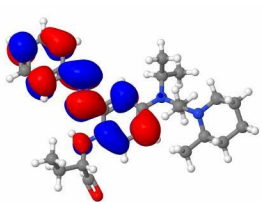
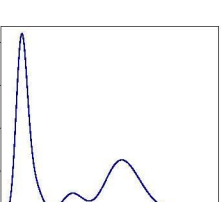
From the plots in Figure 6, it can be seen that PBE0/DGDZVP,

density functional and basis-set combination, generally tends to underpredict the λ_{max} of molecules having an experimental λ_{max} above 520 nm. The average error in prediction using PBE0 functional for molecules with experimental λ_{max} above 520 was 20 nm and it ranged from -60 nm to +15 nm. A majority of the molecules in this group have a high negative prediction error (See ESI). Erroneous fitness measures make it unlikely for potentially promising candidates to be selected for genetic processes, which is a key step in the design process.

From Table 1, it is evident that the best functional to predict the λ_{max} depends on the molecular structure. In about 70 molecules chosen to identify the best DFT functional, it was observed there were many molecules for which, CAM-B3LYP functional gave good estimates of the λ_{max} , while PBE0 and B3LYP functionals were far from the true λ_{max} . Furthermore, there were sets of molecules for which PBE0 or B3LYP functional specifically, gave the best estimates of the λ_{max} . This essentially means, relying on λ_{max} predicted by a fixed DFT functional for all molecules can be spurious. These errors can be minimised by either tuning existing DFT functionals to suit the molecular system being studied or rely on machine learning algorithms such as, partial least squares, support vector machines and random forests, to help predict the best DFT functionals for new molecules obtained from the *de novo* design⁸⁰.

Some of the spurious results observed for the CAM-B3LYP functional may be due to the fact that it does not incorporate a full

Table 2 Structures of new discovered azobenzene molecules using the *de novo* design. Isodensity surfaces of the HOMO and LUMO orbitals (0.02 a.u.) of the structures are shown. TD-DFT spectra plots were made using GaussSum⁷⁹.

Dye	HOMO	LUMO	Absorption Spectra
MOL1 λ_{max} - 636 nm 			
MOL2 λ_{max} - 608 nm 			
MOL3 λ_{max} - 608 nm 			
MOL4 λ_{max} - 595 nm 			
MOL5 λ_{max} - 640 nm 			
MOL6 λ_{max} - 580 nm 			

range separation. The functional has only 65% HF exchange at long range instead of the correct 100% asymptotic HF exchange.

It has been shown⁸¹ that ensuring a 100% asymptotic HF exchange is very important for accurate description of valence ex-

Table 3 Table below summarises lowest energy transitions of all six molecules shown in Table 2. The TD-DFT (B3LYP-DGDZVP) spectra was computed in ethanol solvent. The oscillator strength (f) corresponding to the lowest energy transition, the generation at which the molecule was discovered (Generation) and the transitions are also shown. The first four molecules were discovered when *background knowledge* was given to the EA routine and the last two molecules were identified when starting population was randomly generated by the program.

Molecule	λ_{\max} (nm) [eV]	f	Generation	Major Transitions
MOL1	636 [1.95]	0.2638	Gen001	H→L (98%)
MOL2	608 [2.03]	0.2644	Gen006	H→L (100%)
MOL3	608 [2.03]	0.3399	Gen053	H→L (96%), H→L-1 (2%)
MOL4	595 [2.08]	0.1253	Gen006	H-2→L (11%), H→L (79%)
MOL5	640 [1.94]	0.1682	Gen044	H→L (100%)
MOL6	580 [2.14]	0.1743	Gen083	H→L (100%)

citations in even relatively simple molecular systems. An alternative approach, would be to employ the many-body perturbation theory GW method⁸² in combination with the Bethe-Salpeter equation (BSE) formalism⁸³. Recent results indicate that the GW-BSE method is promising for accurate calculation of the excitation energies of conjugated systems⁸⁴.

A fitness function evaluation that involves TD-DFT and DFT computations, takes considerable time to complete. To speed up the design process, an alternative approach would be to employ quantitative structure-property relationship models (QSPR) to evaluate the fitness^{7,8}. These models although are fast, they tend to be only locally applicable, i.e. fitness estimates using these methods can only be reliable if the molecule being evaluated, is similar to molecules that were used to train the model. Statistical measures⁸⁵ can be used to infer the confidence intervals and help us decide if the model predicted fitness can be trusted.

4 Conclusion

In this work, we have presented a method to design azobenzene based structures with longer absorption wavelengths. A high-throughput DFT and TD-DFT based fitness evaluation algorithm that was integrated with the *de novo* based design strategy. The algorithm was designed to take advantage of the processing capabilities of multiple supercomputing clusters and distributed computing networks to speed up the computations and also provide reliable fitness estimates, which is essential to the design scheme.

Acknowledgements

The Research Council of Norway (NFR) is acknowledged for their financial support eVITA (Grant No. 205273/F30) and the Norwegian center for highperformance computing (NOTUR) for the CPU resources granted for the project. We also thank ChemAxon (<http://www.chemaxon.com>) for free academic use of the Marvin package.

References

- G. Ceder and K. Persson, *Sci. Am.*, 2013, **309**, 36–40.
- J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Roman-Salgado, K. Treppe, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao and A. Aspuru-Guzik, *Energy Environ. Sci.*, 2014, **7**, 698–704.
- M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson and M. Asta, *Scientific Data*, 2015, **2**, 150009.
- M. Korth, *Phys. Chem. Chem. Phys.*, 2014, **16**, 7919–7926.
- X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss and K. A. Persson, *Comp. Mat. Sci.*, 2015, **103**, 56–67.
- V. Venkatasubramanian, K. Chan and J. Caruthers, *Comput. Chem. Eng.*, 1994, **18**, 833–844.
- V. Venkatraman, M. Foscatto, V. R. Jensen and B. K. Alsberg, *J. Mater. Chem.*, 2015, **3**, 9851–9860.
- V. Venkatraman, S. Abburu and B. K. Alsberg, *Phys. Chem. Chem. Phys.*, 2015, **17**, 27672–27682.
- S. Lysgaard, J. S. G. Myrdal, H. A. Hansen and T. Vegge, *Phys. Chem. Chem. Phys.*, 2015, <http://dx.doi.org/10.1039/C5CP00298B>.
- J. B. A. Davis, A. Shayeghi, S. L. Horswell and R. L. Johnston, *Nanoscale*, 2015, **7**, 14032–14038.
- Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen and B. K. Alsberg, *J. Am. Chem. Soc.*, 2012, **134**, 8885–8895.
- M. Hartenfeller, E. Proschak, A. Schüller and G. Schneider, *Chem. Biol. Drug. Des.*, 2008, **72**, 16–26.
- D. Hecht and G. B. Fogel, *J. Chem. Inf. Model.*, 2009, **49**, 1105–1121.
- C. A. Nicolaou and N. Brown, *Drug Discovery Today: Technol.*, 2013, **10**, e427–e435.
- G. Schneider and U. Fechner, *Nat. Rev. Drug Discov.*, 2005, **4**, 649–663.
- F. Dey and A. Cafilisch, *J. Chem. Inf. Model.*, 2008, **48**, 679–690.
- D. Dominique, M.-L. Hélène, L. Gilles and P. Sylvie, *J. Med. Chem.*, 2005, **48**, 2457–2468.
- G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *J. Comput. Chem.*, 1998, **19**, 1639–1662.
- G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.
- J.-M. Yang, *J. Comput. Chem.*, 2004, **25**, 843–857.
- H.-J. Böhm and G. Schneider, *Virtual Screening for Bioactive Molecules*, New York: Wiley-VCH, 2000.
- G. Holger, H. Manfred and K. Gerhard, *J. Mol. Biol.*, 2000, **295**, 337–356.
- M. Rarey, B. Kramer and T. Lengauer, *Bioinformatics*, 1999, **15**, 243–250.
- T. Hansson, J. Marelius and J. Åqvist, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 27–35.
- J. Åqvist, C. Medina and J.-E. Samuelsson, *Protein Eng.*, 1994, **7**, 385–391.
- O. N. Carstensen, J. M. Dieterich and B. Hartke, *Phys. Chem. Chem. Phys.*, 2011, **13**, 2903–2910.
- G. Schneider, *De novo Molecular Design*, Wiley-VCH: Verlag GmbH & Co. KGaA, 2013.
- N. Martsinovich and A. Troisi, *Energy Environ. Sci.*, 2011, **4**, 4473–4495.
- L. Zhang, J. M. Cole, P. G. Waddell, K. S. Low and X. Liu, *ACS Sustainable Chem. Eng.*, 2013, **1**, 1440–1452.
- Z. Wang, Y. Cui, K. Hara, Y. Dan-oh, C. Kasada and A. Shinpo, *Adv. Mater.*, 2007, **19**, 1138–1141.
- B. E. Hardin, H. J. Snaith and M. D. McGehee, *Nat. Photon.*, 2012, **6**, 162–169.
- S. Alex, U. Santhosh and S. Das, *J. Photochem. Photobiol., A*, 2005, **172**, 63–

- 71.
- 33 A. Burke, L. Schmidt-Mende, S. Ito and M. Gratzel, *Chem. Commun.*, 2007, 234–236.
- 34 H. H. Wendy, C. Frederik, M. Hidetoshi and M. P. Laurence, *J. Am. Chem. Soc.*, 2008, **130**, 1367–1375.
- 35 Y. Cui, J. Yu, J. Gao, Z. Wang and G. Qian, *J. Sol-Gel Sci. Technol.*, 2009, **52**, 362–369.
- 36 Q. Peng, Z.-Y. Lu, Y. Huang, M.-G. Xie, S.-H. Han, J.-B. Peng and Y. Cao, *Macromolecules*, 2004, **37**, 260–266.
- 37 Y.-A. Son, S.-Y. Gwon, S.-Y. Lee and S.-H. Kim, *Spectrochim. Acta, Part A*, 2010, **75**, 225 – 229.
- 38 H. M. D. Bandara and S. C. Burdette, *Chem. Soc. Rev.*, 2012, **41**, 1809–1825.
- 39 A. A. Blevins and G. J. Blanchard, *J. Phys. Chem. B*, 2004, **108**, 4962–4968.
- 40 M. Dong, A. Babalhavaeji, S. Samanta, A. A. Beharry and G. A. Woolley, *Acc. Chem. Res.*, 2015, **48**, 2662–2670.
- 41 M. Dong, A. Babalhavaeji, M. J. Hansen, L. Kalman and G. A. Woolley, *Chem. Commun.*, 2015, **51**, 12981–12984.
- 42 S. Samanta, T. M. McCormick, S. K. Schmidt, D. S. Seferos and G. A. Woolley, *Chem. Commun.*, 2013, **49**, 10314–10316.
- 43 A. A. Beharry, O. Sadovski and G. A. Woolley, *J. Am. Chem. Soc.*, 2011, **133**, 19684–19687.
- 44 M. Izquierdo-Serra, M. Gascón-Moya, J. J. Hirtz, S. Pittolo, K. E. Poskanzer, Èric Ferrer, R. Alibés, F. Busqué, R. Yuste, J. Hernando and P. Gorostiza, *J. Am. Chem. Soc.*, 2014, **136**, 8693–8701.
- 45 M. A. Kienzler, A. Reiner, E. Trautman, S. Yoo, D. Trauner and E. Y. Isacoff, *J. Am. Chem. Soc.*, 2013, **135**, 17683–17686.
- 46 H. Nishioka, X. Liang, T. Kato and H. Asanuma, *Angew. Chem. Int. Edit.*, 2012, **51**, 1165–1168.
- 47 J. Garcia-Amoros, A. Sanchez-Ferrer, W. A. Massad, S. Nonell and D. Velasco, *Phys. Chem. Chem. Phys.*, 2010, **12**, 13238–13242.
- 48 N. Brown, B. McKay, F. Gilardoni and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1079–1087.
- 49 F. Herrera and J. L. Verdegay, *Genetic Algorithms and Soft Computing (Studies in Fuzziness and Soft Computing Vol. 8)*, Physica-Verlag Heidelberg, 1996.
- 50 J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.
- 51 L. Wade, *Organic chemistry*, Prentice Hall, 6th edn, 2005.
- 52 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, *ChemMedChem*, 2008, **3**, 1503–1507.
- 53 *Marvin 5.9.3*, 2012, ChemAxon (<http://www.chemaxon.com>).
- 54 S. L. Mayo, B. D. Olafson and W. A. Goddard, *The J. Phys. Chem.*, 1990, **94**, 8897–8909.
- 55 M. J. S. Dewar, E. G. Zebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
- 56 J. J. P. Stewart, *MOPAC2012*, 2012, Stewart Computational Chemistry, Colorado Springs, CO, USA, (<http://OpenMOPAC.net>).
- 57 M. Valiev, E. Bylaska, N. Govind, K. Kowalski, T. Straatsma, H. V. Dam, D. Wang, J. Nieplocha, E. Apra, T. Windus and W. de Jong, *Comput. Phys. Commun.*, 2010, **181**, 1477 – 1489.
- 58 C. Lee, W. Yang and R. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
- 59 G. A. Petersson and M. A. Al-Laham, *The J. Chem. Phys.*, 1991, **94**, 6081–6090.
- 60 D. Jacquemin, J. Preat, E. A. Perpète, D. P. Vercauteren, J.-M. André, I. Ciofini and C. Adamo, *Int. J. Quantum Chem.*, 2011, **111**, 4224–4240.
- 61 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc. Wallingford CT 2009.
- 62 Y. Takano and K. N. Houk, *J. Chem. Theory Comput.*, 2005, **1**, 70–77.
- 63 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
- 64 N. Godbout, D. R. Salahub, J. Andzelm and E. Wimmer, *Can. J. Chem.*, 1992, **70**, 560–571.
- 65 N. O’Boyle, M. Banck, C. James, C. Morley, T. Vandermeersch and G. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
- 66 C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E. Willighagen, *Curr. Pharm. Des.*, 2006, **12**, 2111–2120.
- 67 K. S. Thanthiriatte, and S. R. Gwaltney, *J. Phys. Chem. A*, 2006, **110**, 2434–2439.
- 68 *Notur, Vilje specifications*, Accessed: 09th April 2015, <https://www.notur.no/hardware/vilje>.
- 69 *Notur, Stallo specifications*, Accessed: 09th April 2015, <https://www.notur.no/hardware/stallo>.
- 70 K. Andersson, P. A. Malmqvist, B. O. Roos, A. J. Sadlej and K. Wolinski, *J. Phys. Chem.*, 1990, **94**, 5483–5488.
- 71 J. D. Watts and R. J. Bartlett, *J. Chem. Phys.*, 1994, **101**, 3073–3078.
- 72 E. Runge and E. K. U. Gross, *Phys. Rev. Lett.*, 1984, **52**, 997–1000.
- 73 K. Aidas, J. Kongsted, A. Osted, K. V. Mikkelsen and O. Christiansen, *J. Phys. Chem. A*, 2005, **109**, 8001–8010.
- 74 D. Jacquemin, E. A. Perpète, I. Ciofini and C. Adamo, *Acc. Chem. Res.*, 2009, **42**, 326–334.
- 75 R. Peverati and D. G. Truhlar, *Phil. Trans. R. Soc. A*, 2014, **372**, year.
- 76 C. Adamo and D. Jacquemin, *Chem. Soc. Rev.*, 2013, **42**, 845–856.
- 77 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51 – 57.
- 78 H. Moustroph and J. Epperlein, *J. f. prakt. Chemie. Band*, 1981, **5**, 755–775.
- 79 N. M. O’boyle, A. L. Tenderholt and K. M. Langner, *J. Comput. Chem.*, 2008, **29**, 839–845.
- 80 V. Venkatraman, S. Abburu and B. K. Alsberg, *Chemom. Intell. Lab. Syst.*, 2015, **142**, 87–94.
- 81 B. M. Wong and J. G. Cordaro, *J. Chem. Phys.*, 2008, **129**, 214703.
- 82 F. Aryasetiawan and O. Gunnarsson, *Rep. Prog. Phys.*, 1998, **61**, 237.
- 83 E. E. Salpeter and H. A. Bethe, *Phys. Rev.*, 1951, **84**, 1232–1242.
- 84 B. Baumeier, D. Andrienko, Y. Ma and M. Rohlfing, *J. Chem. Theory Comput.*, 2012, **8**, 997–1002.
- 85 P. Carrió, M. Pinto, G. Ecker, F. Sanz and M. Pastor, *J. Chem. Inf. Model.*, 2014, **54**, 1500–1511.