
Article:
Shaw, R.A. orcid.org/0000-0002-9977-0835, Manian, A., Lyskov, I. et al. (1 more author) (2021) Efficient enumeration of bosonic configurations with applications to the calculation of non-radiative rates. The Journal of Chemical Physics, 154 (8). 084102. ISSN 0021-9606

https://doi.org/10.1063/5.0039532

Efficient enumeration of bosonic configurations with applications to the calculation of non-radiative rates

Robert A. Shaw, Anjay Manian, Igor Lyskov, and Salvy P. Russo[a)]

ARC Centre of Excellence in Exciton Science, School of Science, RMIT University, Melbourne, VIC 3000, Australia

(Dated: 20 January 2021)

This work presents algorithms for the efficient enumeration of configuration spaces following Boltzmann-like statistics, with example applications to the calculation of non-radiative rates, and an open-source implementation. Configuration spaces are found in several areas of physics, in particular wherever there are energy levels that possess variable occupations. In bosonic systems, where there are no upper limits on the occupation of each level, enumeration of all possible configurations is an exceptionally hard problem. We look at the case where the levels need to be filled to satisfy an energy criterion, for example a target excitation energy, which is a type of knapsack problem as found in combinatorics. We present analyses of the density of configuration spaces in arbitrary dimensions, and how particular forms of kernel can be used to envelope the important regions. In this way, we arrive at three new algorithms for enumeration of such spaces that are several orders of magnitude more efficient than the naive brute force approach. Finally, we show how these can be applied to the particular case of internal conversion rates in a selection of molecules, and discuss how a stochastic approach can in principle reduce the computational complexity to polynomial time.

[a)]Electronic mail: salvy.russo@rmit.edu.au

I. INTRODUCTION

There are several problems in chemical physics where one needs to enumerate points in an occupation or configuration space, subject to some criterion on those configurations. Important examples arise in statistical thermodynamics, where the calculation of partial partition functions over some subset of microstates is necessary for determination of thermodynamical constants.[1] Similar quantum-statistical principles find applications ranging from path-integral molecular dynamics[2] and configuration interaction[3,4] methods for bosons, with enumeration of configurations being a key difficulty.[5,6] The focus of this work will be that a number of molecular electronic properties can be determined from knowledge of the configuration of quanta in vibronic modes subject to some energy constraint on the configuration.[7–10] In combinatorial mathematics, the problem of selecting integer occupations to satisfy a total "weight" is known as the knapsack problem.[11] In the simplest version of this, you have a knapsack that can carry a fixed volume of objects, and you have a selection of objects to fill it with, each of which has an inherent volume and value. The problem is to fill the knapsack such as to maximise the total value, while not exceeding the total possible volume. In our physical equivalent in this paper, the "knapsack" is the occupation vector, with each mode possessing an energy (equivalent to volume) and Franck-Condon factor[12] (equivalent to the value). We then not only wish to find the configuration that gives the best total value, but also to enumerate all possible configurations that satisfy the volume (energy) requirement.

The difficulty with this problem is that it is of non-polynomial complexity,[13] and as such no polynomial-time algorithm is known that can guarantee the correct solution. That is not to say that solutions do not exist; an obvious route would be to simply enumerate every possible configuration, and evaluate its volume and value, selecting the best possible solution from these. This has the added advantage that it solves our extended problem of determining all the configurations within the energy crtierion. However, the number of configurations increases exponentially with the number of modes, rapidly making use of such a brute-force algorithm intractable. Specifically in the 0-1 knapsack problem (where each object or mode is either included or not), the number of configurations follows 2^M where M is the number of modes. The best known heuristic algorithm (full polynomial-time approximation scheme) to solve the 0-1 problem scales as $\mathcal{O}(\epsilon^{-1}M^{1/\epsilon})$, to get a solution with value at worst $(1-\epsilon)$ times the optimum.[14] This does not solve the wider enumeration problem, however, and by

necessity approaches the factorial scaling as the threshold ϵ is tightened. Note that the more general bosonic problem, where each mode can have any integer occupation, can always be rewritten in terms of the 0-1 problem by allowing copies of the modes.

While the algorithms we present here are more widely applicable, we will focus on the specific problem of determining non-radiative rates. Understanding molecular photophysical processes is an important and difficult problem. This is particularly true in the field of exciton science,[15,16] and in the design of optical devices and light-harvesting materials.[17–19] Typically, the goal is to either maximize or minimize the photoluminescent quantum yield of such devices, determined as the ratio of the radiative decay rate to the sum of radiative and non-radiative rates. The former comprise fluorescence and phosphorescence, while the latter are predominantly internal conversion (IC, spin preserving) and intersystem crossing (ISC, spin flipping). These non-radiative processes are facilitated largely through coupling of electronic and vibrational states, and thus occur when the molecule is excited into a configuration where the vibronic quanta satisfy an energy criterion. The simplest example would be excitation from the ground to first excited electronic singlet states. The vibrational manifolds of each state then allow for an energy window of possible configurations that can result in internal conversion occurring, and each such configuration can then be weighted by a probability of resulting in either IC or fluorescence. As such, the theoretical determination of the rate is naturally formulated as an example of the knapsack problem.

Previous work by Valiev *et al.*[20–22] has focused on the solution of this problem using the Plotnikov, Robinson, and Jortner (PRJ) formalism for non-radiative energy transfer.[9,10] In their original and more recent papers, they give a scheme for the determination of the various quantum chemical properties that need to be calculated in determining these rates, but focus more on the application of the method than on algorithmic development for the exploration of configuration space. In this work, we will first briefly recap the theory underlying the formalism, before analysing the density of the configuration space. The main focus will then be on devising efficient algorithms for the full characterisation of this space, both within this specific case study, but also more generally for other knapsack-type physical problems. Finally we will demonstrate the effectiveness of these algorithms, and provide a fully open-source implementation which can be found at[23].

II. THEORY

A. Background

For a transition between two electronic states, $|i\rangle$ (initial) and $|f\rangle$ (final), with energy difference E_{if}, the general non-radiative rate in the PRJ formalism is written as[20]:

$$k_{\mathrm{nr}} = \frac{4}{\Gamma_f} \sum_{\mathbf{n}}^{E(\mathbf{n})=E_{if}} |V_{if}(\mathbf{n})|^2 \tag{1}$$

where the sum is over vibronic configurations $\mathbf{n} = (n_1, n_2, \ldots, n_M)$, Γ_f is the relaxation width of state f, and V_{if} is a coupling potential that depends on which rate is being calculated. In this, we have assumed that the relaxation width does not depend strongly on $\mathbf{n}$, and that $E_{if} \ll \Gamma_f$. These conditions generally hold true for temperatures around and below room temperature,[24] typically matching experimental conditions.

Each $\mathbf{n}$ has an associated Franck-Condon factor, determined using the Huang-Rhys (HR) factors, y_j, of the M vibrational modes:

$$\mathrm{FC}(\mathbf{n}) = \prod_{j=1}^{M} \left(\frac{e^{-y_j} y_j^{n_j}}{n_j!} \right) \tag{2}$$

This factor effectively determines the extent of the vibrational overlap, and thus the strength of the contribution to the rate, of a configuration. From this we see that modes increase with increasing y_j and decrease with increasing n_j. That is, modes with large HR factors can in general have larger occupations and still give significant contributions, or conversely, those with small y_j are more likely to have low occupations. This will be important later, as it suggests a way to assess the importance of a configuration.

There are then two physical regimes in which the PRJ formalism can be applied: under the Franck-Condon approximation, or the Herzberg-Teller approximation.[21] For internal conversion, which we will be focusing on here, these two formulations are written as follows.

The formulas for intersystem crossing are similar and can be found in ref.[22].

$$V_{if}^{IC,FC}(\mathbf{n}) = -\sum_{j=1}^{M}\left(\sum_{\nu q} m_{\nu}^{-1}\,\langle i|\frac{\partial}{\partial R_{q\nu}}|f\rangle\, B_{\nu qj}\right)\left(\frac{\omega_j(n_j-y_j)^2}{2y_j}\cdot \mathrm{FC}(\mathbf{n})\right)^{1/2} \tag{3}$$

$$\begin{aligned} V_{if}^{IC,HT}(\mathbf{n}) = &-\sum_{j=1}^{M}\sum_{j'=1}^{M}\left(\sum_{\nu q}\sum_{\nu' q'}(m_{\nu}m_{\nu'})^{-1}\,\langle i|\frac{\partial^2}{\partial R_{q\nu}\partial R_{q'\nu'}}|f\rangle\, B_{\nu qj}B_{\nu' q' j'}\right) \\ &\times\left(\frac{(n_{j'}+y_{j'})^2}{2\omega_{j'}y_{j'}}\right)^{1/2}\left(\frac{\omega_j(n_j-y_j)^2}{2y_j}\cdot \mathrm{FC}(\mathbf{n})\right)^{1/2} \end{aligned} \tag{4}$$

where ω_j is the energy of the jth mode, and $R_{\nu q}$ is the qth coordinate of the νth atom with mass m_ν. Additionally, these depend on nuclear gradients along the vibrational modes, $B_{\nu qj}$, and vibronic couplings between the initial and final electronic states. Writing them in this way demonstrates how the rates can easily be simplified into dot products or matrix multiplications of a part that depends exclusively on the coordinates, and a part that depends on the choice of configuration:

$$k_{IC,FC} = \frac{4}{\Gamma_f}\sum_{\mathbf{n}}^{E(\mathbf{n})=E_{if}}\left[\mathbf{a}_{\mathrm{FC}}(\mathbf{R})\cdot\mathbf{z}_{\mathrm{FC}}(\mathbf{n},\mathbf{y})\right]^2 \tag{5}$$

$$k_{IC,HT} = \frac{4}{\Gamma_f}\sum_{\mathbf{n}}^{E(\mathbf{n})=E_{if}}\left[\mathbf{z}_{\mathrm{HT}}(\mathbf{n},\mathbf{y})\cdot\mathbf{A}_{\mathrm{HT}}(\mathbf{R})\cdot\mathbf{z}_{\mathrm{FC}}(\mathbf{n},\mathbf{y})\right]^2 \tag{6}$$

where the $\mathbf{z}$, $\mathbf{a}$, and $\mathbf{A}$ tensors are defined implicitly in equations 3 and 4.

Most importantly all the rates, including the ISC ones, are modulated through a dependence on $\mathrm{FC}(\mathbf{n})$. Moreover, they are formed as a sum over configurations satisfying a fixed energy criterion, leading to our connection to the knapsack problem. Writing it as $E(\mathbf{n}) = E_{if}$ is somewhat disingenuous, however. For a system with a finite number of modes of fixed energy, the probability of finding a configuration (where occupations are necessarily integers) is infinitesimally small; this is due to the set of integers being countably infinite, while the set of real numbers (i.e. possible values of E_{if}) is uncountably infinite. Physically, asserting an exact energy criterion would be nonsensical anyway - the excitation is occurring between two electronic states with vibrational manifolds. Even at very low temperatures, the excitation band will have non-zero width, implying that in reality the energy criterion is

$$|E(\mathbf{n}) - E_{if}| \leq \delta \tag{7}$$

where δ is some energy window reflecting the thermal variation in acceptable excitation

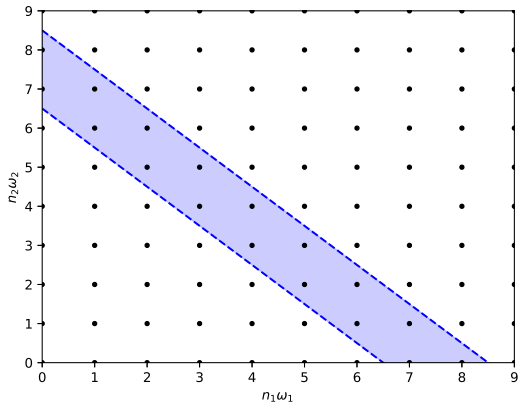

FIG. 1. A representation of the enumeration problem when $M = 2$. The configurations are the lattice points, shown as bold black dots, while the energy criteria, $E = E_{if} \pm \delta$, are described by the dashed blue lines. The 'acceptable' region is then shaded light blue, showing the density of configurations. As M increases, the size of this volume will also increase.

energies. To determine the rates, we thus need to enumerate configuration space within this energy window.

B. Density of configuration space

Each configuration vector, $\mathbf{n}$, can be thought of as a point in an M-dimensional configuration space, where each axis is rescaled by the weight or volume, w_j, assigned to that object. The total volume of the configuration is thus the sum $w_j n_j$. The fixed-volume criterion then describes an $(M-1)$-dimensional plane, and the expanded knapsack problem becomes finding all the lattice points that lie on that plane. If we extend to the range in equation 7, this describes a volume in configuration space bounded by two such planes, and we wish to find all the lattice points within that volume. Intuitively, as M increases, the number of possible lattice points in either instance will increase too, and this will increase further if we use a wider window. In this section, we will look more rigorously at how many significant configurations there are in such a system, and how the density of configurations behaves asymptotically. This will allow us to assess how successful any approximate methodologies are at characterizing the space.

To determine the behaviour, we consider how the volume of the acceptable region, and

the number of lattice points within that region, increases with both M and the window δ. The simplest, two-dimensional case is shown in Figure 1, where the hyper-surface is simply a line described by $\omega_2 n_2 = E \pm \delta - \omega_1 n_1$. The acceptable volume is then the difference between the larger triangle and the smaller triangle. If we were to expand to three dimensions, these would be octants of a tetrahedron, and more generally in M dimensions, a pair of regular M-simplexes. The volume of such a simplex with side length L is $L^M/M!$. By defining our energy scale as $E + \delta \equiv 1$, the larger simplex has a volume of $1/M!$, and thus the acceptable region as a proportion of this total volume is:

$$v_M(\delta)/V_M = \frac{1 - (1 - 2\delta)^M}{M!} \cdot M! = 1 - (1 - 2\delta)^M \tag{8}$$

Now we consider the total number of lattice points within the larger simplex. If we consider the 2D case once more, then the number of lattice points in the square with side $E + \delta$ is simply the product of $1 + n_{1,\max}$ and $1 + n_{2,\max}$. Similarly, in higher dimensions, the hypercube contains the product of all such $1 + n_{i,\max}$. These maximum values are easily computed as

$$n_{i,\max} = \left\lfloor \frac{E + \delta}{\omega_i} \right\rfloor \tag{9}$$

The M-simplex is then $1/M!$ of the total volume of the hypercube,[25] and so on average contain

$$\frac{1}{M!} \prod_{i=1}^{M} (1 + n_{i,\max}) \sim \frac{(E + \delta)^M}{M! \prod_i \omega_i}$$

such points. Finally, combined with equation 8, the acceptable region must on average contain

$$\tilde{N}_M(\delta) \sim \frac{E^M}{M! \prod_i \omega_i} \left[1 - (1 - 2\delta)^M\right] \cdot \left(1 + \frac{\delta}{E}\right)^M \tag{10}$$

For $\delta \ll E$, we can expand the two terms in brackets as Taylor series. Retaining only the term linear in δ, this yields

$$\tilde{N}_M(\delta) \sim \frac{2E^M \delta}{(M-1)! \prod_i \omega_i} + \mathcal{O}(\delta^2)$$

showing that the number of lattice points asymptotically tends to zero as the window shrinks, as expected. As mentioned earlier, the probability of a lattice point lying exactly on the hypersurface $E(\mathbf{n}) = E$ is vanishingly small. In the regime of large M, we instead expand the brackets as binomial series. Approximating each weight as some fraction, $\omega \approx \alpha E/M$,

we get

$$\tilde{N}_M(\delta) \sim 2\left(\frac{M}{\alpha}\right)^M \sum_{k=1}^{M}\sum_{l=0}^{M} \frac{(-2)^{k-1}M!}{(M-k)!(M-l)!k!l!}\delta^{k+l}E^{-l}$$

Given the assumption again that $\delta \ll E$, this is dominated by the first term in the double sum, i.e. $k = l = 0$, This gives the much simpler asymptotic expression:

$$\tilde{N}_M \sim \frac{2M^M}{\alpha^M M!} \sim 2\left(\frac{e}{\alpha}\right)^M \tag{11}$$

where we have used Stirling's approximation. From this we see that for a fixed value of δ, the number of lattice points explodes exponentially in higher dimensions. Any approximations, therefore, that rely on finding a single acceptable point in the allowable region will therefore give increasingly worse results as the number of dimensions increases. Efficient strategies for finding only the most important configurations, using knowledge of the "values" of each point, are thus necessary to make this problem tractable.

C. Franck-Condon weightings

Having determined what the density of the configuration space is, the next step is to characterise the importance of various regions in that space. This will be problem specific, depending on how we have defined the values in the knapsack problem. In the case of non-radiative rates, we want to find the configurations that give the largest contributions to the rate. As these are mediated primarily through Franck-Condon factors in all cases, we wish to find all configurations that have a factor within some threshold. We can rewrite the product in equation 2 as a sum in the exponent to get:

$$\mathrm{FC}(\mathbf{n}) = \exp\left\{\sum_j \left(n_j \ln y_j - y_j\right)\right\} \cdot \prod_{j=1}^{M} \left(n_j!\right)^{-1}$$

All of the lattice points in a region can be found as the union of all the lattice points on the planes defined by fixed $|n|$ that intersect the region, where $|n| = \sum_j n_j$ is the total occupation number for that configuration. If we estimate the individual n_j as their average values, $|n|/M$, and y_j by the average value, $\bar{y}$, we can use Stirling's approximation to get

$$\mathrm{FC}(\mathbf{n}) \sim \left(\frac{M}{2\pi|n|}\right)^{|n|+M/2} (2\pi e\bar{y})^{|n|} \exp\left(-M\bar{y}\right) \tag{12}$$

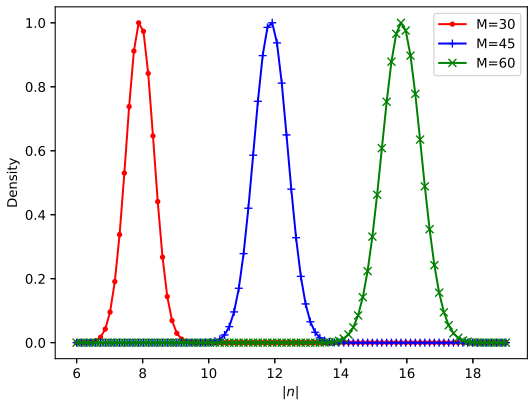

FIG. 2. Normalized importance factor for the Franck-Condon kernel ($\bar{y} = 1$) as a function of the total occupation, $|n|$, for various sized dimensions, M.

We note that any weighting function (i.e. probability distribution) defining the value of the objects in the knapsack problem will have this asymptotic form of envelope if its kernel follows the functional form

$$\rho(\mathbf{n}) \sim p_1(|n|) \cdot \exp\left(p_2(|n|)\right) \cdot \prod_{j=1}^{M} (n_j!)^{-1} \tag{13}$$

where p_1 and p_2 are arbitrary polynomials with real coefficients. This encompasses many different probability distributions, including those generated by partition functions in various ensembles,[1] or in general where Boltzmann-like statistics are present.[26]

The importance of a given value of $|n|$ is then equation 13 weighted by the number of acceptable configurations with that $|n|$. As will be discussed in more detail later, the number of such configurations follows a binomial distribution, from some $|n|_{\min}$ to $|n|_{\max}$, which are determined by the particular choice of target energy. By approximating this symmetric binomial distribution as a normal distribution, the proportion, $p(|n|)$, of total acceptable configurations with fixed $|n|$ is given by

$$p(|n|) \sim \sqrt{\frac{\alpha E}{M\pi\delta}} \exp\left\{-\frac{E}{M\alpha\delta}\left(\alpha|n| - M\right)^2\right\}$$

where α is as defined earlier. Using the result from equation 11, we therefore have that the weighted importance value, $P(|n|)$, is given by

$$P(|n|) = p(|n|) \cdot N_M \cdot \rho(|n|) \sim p(|n|) \left(\frac{|n|}{M}\right)^{M/2-|n|} \exp(-kM) \tag{14}$$

where the constant k is dependent on the kernel; for the Franck-Condon factors, $k = \bar{y}$. From this we see that the importance factor follows a form of Gamma distribution, with the center of the distributing shifting higher with increasing dimension, M, and with decreasing k.

Figure 2 shows examples of these distributions for various values of M and $|n|$, with $y_{\min}$ set at 0.5. This analysis shows that the FC factor acts as an envelope on configuration space, greatly reducing the number of configurations that need to be considered, as those configurations outside the main envelope will not contribute significantly to the rate. From equation 12, we see that the width of the distribution follow $M/2$, implying that the number of significant $|n|$ increases linearly with dimension. However, this is also mitigated by the exponential, which adds a factor of e^{-y} for every extra mode, which is necessarily less than unity, as $y > 0$. Therefore, theoretically, the problem is not of factorial complexity, as first seemed.

III. METHODS

The general problem we are trying to solve is to find all configurations $\mathbf{n}$ that satisfy the energy criterion in equation 7, that have values, $\rho(\mathbf{n})$, greater than some given threshold. The only assumptions that we make are that the occupations, n_j, are necessarily integers, and that the "energy" can be written as $\sum_j n_j w_j$ for positive weights, w_j. We have shown some of the theoretical properties of such a system in the previous section, and in the present section, we will present algorithms for finding the solution. Implementations of these algorithms can be found in the open-source Knapsack software package.[23]

A. Screened brute force approach

In the simplest, brute-force approach to solving the problem, we enumerate every possible configuration and test it against the energy criterion. In this manner, no significant configurations can be missed. It may at first seem that there are infinitely many such configurations, but we can place an upper bound on the occupation of each mode as no $n_j w_j$ can be greater than the maximum allowed energy, $E + \delta$. As such equation 9 gives upper bounds for each occupation.

There are then two possible algorithmic approaches to enumerating these for systems of arbitrary dimensions. The first is through hashing, where we loop an index, $i(\mathbf{n})$ from 1 to $N = \prod_j n_{j,\max}$, where this index corresponds to a configuration. The configuration can be reconstructed as

$$i(\mathbf{n}) = n_1 + \sum_{j=2}^{M} n_j \left(\prod_{k=1}^{j-1} n_{k,\max} \right)$$

This is computationally efficient, because you are simply increasing a counter, without the need for constructing M nested loops.

Algorithm 1 Screened brute-force algorithm for finding configurations, $\mathbf{n}$, that satisfy the energy criterion, equation 7, within a threshold T on the value $\rho(\mathbf{n})$.

For each integer $0 \leq t \leq -\log_{10} T$ and mode j, tabulate $\mathcal{N}(j,t) = n_{j,\max}$ such that

$$\rho(0, \ldots, n_{j,\max}, \ldots, 0) \geq 10^{-t}$$

For each possible value of n_j, tabulate $\mathcal{T}(j,n) = t$ such that $\rho(0, \ldots, n_j = n, \ldots, 0) \geq 10^{-t}$

Call ITERATE($j = 1, \mathbf{n} = (\mathcal{N}(1, t_{\max}), 0, 0, \ldots)$

procedure ITERATE($j, \mathbf{n}$)

    if $j = M$ **then**

        for $i \leftarrow 0, n_j$ **do**

            Let $\mathbf{n}' = \mathbf{n}$ but with $n_j = i$

            Check $E(\mathbf{n}')$ against criterion

        end for

    else

        for $i \leftarrow 0, n_j$ **do**

            Let $\mathbf{n}' = \mathbf{n}$ but with $n'_j = i$

            Set t to the maximum of 0 or

$$t_{\max} - \sum_{k<j+1} \mathcal{T}(k, n'_k)$$

            Set $n'_{j+1} = \mathcal{N}(j+1, t)$

            Call ITERATE($j + 1, \mathbf{n}'$)

        end for

    end if

end procedure

However, we can greatly improve the efficiency of the screening if we consider that $n_{j,\max}$ should change dependent on the occupations for all $k < j$. That is to say that if the threshold is T, then given the occupation n_1, $n_{2,\max}$ should be calculated with a threshold of $T/\rho(\mathbf{n})_1$, and so on. The efficiency of this will then be affected by our choice of ordering of the modes, but this can be optimized by ordering them beforehand such that the modes that give the strongest contributions to ρ are first. In the case of the FC factors, this equates to ordering by decreasing magnitude of the Huang-Rhys factors. We cannot use the index hashing approach with this kind of adaptive screening, though, because the indexing no longer follows the simple form given above. In fact, even determining the exact number of configurations, N, is a problem with the same computational complexity as finding all the configurations themselves.

Instead we can use a recursive approach, combined with pre-tabulation of the maximum values of n_j for each threshold up to the minimum threshold, T. This screened brute-force algorithm is described in Algorithm 1. The main downfall of this approach is that compilers have hard limits on the level of recursion allowed, meaning that there is a fundamental limit on the maximum possible number of modes, M. Additionally, recursion, especially at high depths of recursion, can be notably slower than the hashing approach outlined earlier. The efficiency of this approach therefore needs to come from heavy screening, effectively based on limiting total occupation. We can estimate the total number of configurations that will be enumerated in this manner as follows, for the example where $\rho(\mathbf{n})$ is the Franck-Condon factor.

Taking logarithms of equation 2 and rearranging we see that the maximum n for a given threshold $T = 10^{-t} \equiv e^{-\tilde{t}}$ can be estimated from

$$y_j - \tilde{t} = \tilde{n}_{j,\max} \ln y_j - \ln\left(\tilde{n}_{j,\max}!\right) \approx \tilde{n}_{j,\max}\left[\ln y_j - \ln \tilde{n}_{j,\max} + 1\right]$$

where we have again used Stirling's approximation. Rearranging and assuming a general maximum n of around 20, such that $\ln \tilde{n} < 3 \approx \ln 20$, we get

$$\tilde{n}_{j,\max}(\tilde{t}) < \frac{\tilde{t} - y_j}{2 - \ln y_j} \tag{15}$$

Therefore, following Algorithm 1, the total number of configurations can be estimated as

$$\tilde{N} = \sum_{n_1=0}^{\tilde{n}_{1,\max}(\tilde{t})} \sum_{n_2=0}^{\tilde{n}_{2,\max}(\tilde{t}_1)} \cdots \sum_{n_{M-1}=0}^{\tilde{n}_{M-1,\max}(\tilde{t}_{M-2})} \tilde{n}_{M,\max}(\tilde{t}_{M-1})$$

where the conditional threshold $\tilde{t}_j$ is defined as $\tilde{t} - \log_{10} \mathrm{FC}(n_1, n_2, \ldots, n_{j-1}, 0, \ldots)$. This is seemingly a very complex sum, and we leave the somewhat involved algebraic manipulations to the supplementary material. However, the result is fairly simple:

$$\tilde{N} \sim \bar{n}^{M-1} - (1 + \ln \bar{y} + \ln \bar{n})\bar{n}^2 + \mathcal{O}(\bar{n}) \tag{16}$$

where $\bar{n}$ is the maximum possible n_j and $\bar{y}$ is the minimum y_j. While this is clearly greatly reduced from the $\bar{n}^M$ scaling of the unscreened method, it is still exponential, and will become unfeasible for large M.

B. Reduction to quasi-polynomial time

As a result, we want to find a way to reduce the scaling to something computationally feasible. The key to this is the enveloping noted earlier in equation 12. This suggests that we can select a range of $|n|$ such that ignoring all configurations with a total occupation outside this range can be ignored without affecting the calculated value. The number of configurations $N_M(|n|)$ with a fixed $|n|$ is given by

$$N_M(|n|) = \frac{1}{(M-1)!} \prod_{k=0}^{M-2} (|n| + k) \tag{17}$$

That is, the leading term goes as $|n|^{M-2}/(M-1)!$. As we will show shortly, this is strictly still asymptotically exponential in M, but the mantissa is close enough to unity that, for a pragmatic range of M, the complexity appears to behave polynomially.

To see this, we note that for a fixed value of $|n|$, the leading term goes to zero as M goes to infinity, and has a maximum at approximately $|n| = M - 1$. The question then becomes how does the modal value of $|n|$ depend on M. Differentiating equation 14 with respect to $|n|$, and writing $x = |n|/M$, the $|n|_{\mathrm{mode}}$ is found at the solution of

$$\frac{2E\alpha}{\delta} x + \ln x - \frac{1}{2x} = \frac{2E}{\delta} - 1 \tag{18}$$

We show numerical solutions for this for varying values of α and $2E/\delta$ in Figure 3. Notably - and this is shown rigorously in the supplementary material - there is a limiting solution for $2E/\delta$ sufficiently large. That is, for an energy window less than or equal to 2% of the total energy, the modal $|n|$ is given by $|n| = M/\alpha$. Equation 17 then tells us that the number of configurations asymptotically follows $(e/\alpha)^M$, which is exponentially increasing

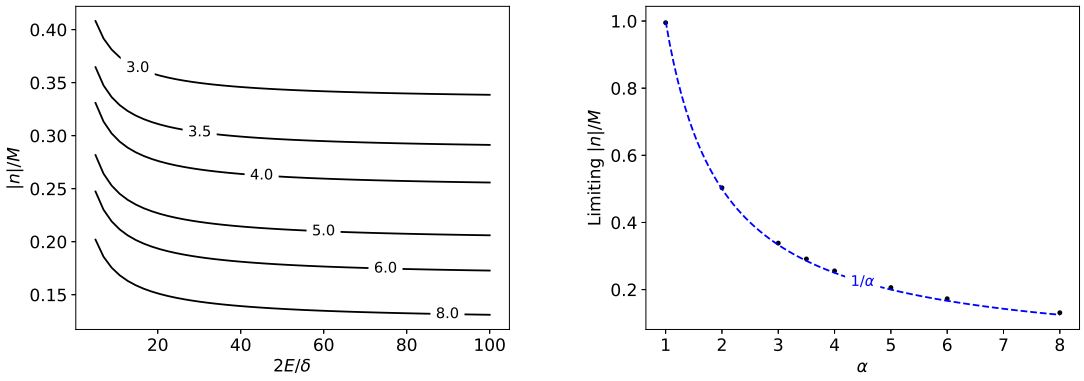

FIG. 3. The left hand panel shows numerical solutions of equation for $x = |n|/M$ in equation 18 for various values of α (marked on the lines) as a function of the energy parameters. These all tend to an asymptote in the large-E/δ limit, and these are plotted as a function of α in the right hand panel, with the curve $x = 1/\alpha$ overlaid.

for $\alpha < e$, and decreasing for $\alpha > e$. The value of α is essentially a measure of how much each individual mode contributes to the total energy, on average. The behaviour will therefore depend heavily on the spectrum of ω values: spectra with large spacings will have larger values of α, leading to a reduced number of configurations, while very densely packed spectra will have small α and much greater numbers of configurations. We will see later that the value of α in the case of the Franck-Condon factors is usually very close to, but slightly less than, e, and as such the scaling appears locally to be polynomial, even though it is strictly speaking exponential; as such, we refer to it as 'quasi-polynomial'.

The algorithm for following such a procedure is also easily adapted from Algorithm 1, and can make use of the same screening, yielding similar $\mathcal{O}(|n|^2 \ln |n|)$ time savings. Only two changes are needed. We add an argument $|n|$ to the ITERATE procedure and in line 6, we replace $i \leftarrow 0, n_j$ with

$$i \leftarrow \max(|n|, 0), \min(|n|, n_M)$$

Then, in line 15 where the recursion happens, we pass the new argument $|n| - i$.

All we require, then, is a method of estimating the minimum and maximum values of $|n|$ required to give values within a given threshold. Alternatively, we can try to estimate these by finding the $|n|$ for which $\rho(\mathbf{n})$ is at a maximum, and estimate the spread of the Gamma distribution. For the former, we note that equation 17 tells us that, agreeing with intuition, the number of configurations increases with $|n|$, and so it is much more important to find a

tight bound on the maximum $|n|$ than for the minimum $|n|$. In this regard, we can trivially compute a lower bound via the minimum target energy as:

$$|n|_{\min} = \left\lceil \frac{E - \delta}{\omega_{\min}} \right\rceil \tag{19}$$

For the upper bound, we must analyse $\rho(\mathbf{n})$. For the specific case of the Franck-Condon factor, from equation 2 that for a given threshold, $T = 10^{-t}$, we have that, approximately,

$$10^{-t} < \exp\left(|n| \ln y_{\max} - \sum_j y_j \right) \left(\frac{M}{|n|} \right)^M$$

where we have estimated the individual n_j as $|n|/M$ and used the fact that $1/n! < 1/n$. Under the reasonable assumption that $y_{\max}$ has no dependence on M, this rearranges to give the estimate

$$|n|_{\max} = M \cdot \exp\left\{ \left(t \ln 10 - \sum_j y_j \right) / M \right\} \tag{20}$$

As expected from earlier, this increases with M, but is also somewhat affected by the values of y_j. Interestingly, if we write the sum in the exponential as $\sum_j y_j = M\bar{y}$ where $\bar{y}$ is the average value of y_j, we can expand the exponential as a Taylor series to see that for large M:

$$|n|_{\max} \sim M - k + \mathcal{O}\left(\frac{1}{M} \right)$$

That is, the maximum increases roughly linearly with M. This, along with the analysis of the spread of the distribution of ρ from earlier (which is also linear in M), means that if we can determine approximately the modal $|n|$, the scaling from equation 17 will be considerably reduced.

To do this, we use the method of Lagrange multipliers. In their most recent paper on the topic,[22] Valiev and coworkers demonstrated that they were using a very similar method to estimate the total contribution from the Franck-Condon factors. As a result, we are able to replicate their approximations directly for the calculation of non-radiative rates, and we will discuss in the results sections how our analysis here demonstrates that it becomes an increasingly poor approximation as M increases. The method of Lagrangian multipliers is well-known and we do not need to discuss it in any detail. We wish to find the maximum of the distribution, $\rho(\mathbf{n})$, which we assume to have kernel of the form given in equation 13. This is subject to the energy constraint in equation 7. The Lagrangian is thus

$$\mathcal{L}(\mathbf{n}) = \ln \rho(\mathbf{n}) - \lambda \left(\sum_j n_j \omega_j - E \right)$$

where λ is a Lagrange multiplier. Differentiating with respect to n_j and rearranging gives

$$n_j \exp\left\{-\frac{p_1'(n_j)}{p_1(n_j)} - p_2'(n_j)\right\} = \exp\left(-\lambda\omega_j\right) \tag{21}$$

which can then be solved for n_j and summed to give the modal value of $|n|$. The multiplier can be found by standard optimization means, by finding the value of λ such that equation 21 satisfies the energy criterion.

The underlying assumption of this approach is that we are allowing the n_j to take non-integer values, as otherwise the Lagrangian is discontinuous and thus undifferentiable. The resulting calculated n_j will be fractional. Therein lies the problem with using this as a method for actually evaluating ρ - the configuration generated is completely invalid, and you are effectively approximating a sum across a distribution with its value at the maximum. As we know from the analysis earlier, the density of configurations increases exponentially with M for nonzero δ, and as such, this type of estimate will be very poor for large M. On the other hand, it will also likely be poor for small M, because the nearest valid configurations - of which there will be relatively few - will not achieve this maximum value, and thus the maximum will be a considerable overestimate. However, for our purposes, it is an excellent manner for determining the two nearest integer $|n|$ to the mode of the distribution, and the value of ρ at that point. Combined with our estimate for the maximum, we can then interpolate the rate at which the distribution decays by using the functional form in equation 13, effectively making the number of fixed-$|n|$ values constant. We will therefore refer to this as the fixed-$|n|$ algorithm.

C. Stochastic sampling of configurations

Another way to arrive at the conclusion of equation 18, but for general ρ of the form in equation 13 is to consider the following. If the spread of ρ follows $M/2$ and $|n|_{\max}$ follows M, as per the previous analysis, we crudely expect the most important $|n|$ to be at around $M/2$. The fixed-$|n|$ algorithm from equation 17 will then follow the *quasi*-polynomial scaling; it is still strictly asymptotically exponential, but the asymptotics will only apply for relatively large M. If we wish to go to even larger systems, however, we need some non-deterministic or heuristic method of characterising configuration space.

The nature of the problem suggests that we can very simply stochastically sample the space by randomly selecting configurations. That is, for each j, we randomly select some

n_j between 0 and $n_{j,\max}$, then test the overall configuration to see if it satisfies the energy criterion. Such a method would naturally scale linearly with M for a fixed number of samples, as there are M random numbers generated per sample. However, the number of samples necessary for such a uniform prior would necessarily scale factorially with M, making it a largely pointless endeavour. If we can find a starting guess in the manifold of acceptable configurations to seed the sampling, we could restrict the space that needs to be explored by only allowing samples of n_j close to the guess. The natural manner to do this is to set a maximum number of occupations to change and/or the maximum amount that $|n|$ is allowed to change. If a configuration so generated is 'acceptable', it is added to the pool of guesses, from which the next sample is generated.

However, from equation 11, we know that even this much reduced space scales exponentially with M. We can follow the deterministic approach of the previous sections, applying the distribution ρ as the prior to our sampling procedure. We attach a weight to allowing a mode to increment or decrement by one based on the change to the 'value', as defined by ρ:

$$p_j^{\pm} = \ln\left\{\frac{\rho(\mathbf{n} : n_j)}{\rho(\mathbf{n} : n_j \pm 1)}\right\} \tag{22}$$

We take logarithms under the assumption that ρ follows the kernel in equation 13 and as such is exponential. The probability of the mode being selected is then simply

$$P(n_j \to n_j \pm 1) = \frac{p_j^{\pm}}{\sum_j p_j^{\pm}} \tag{23}$$

Equation 23 is certainly not the only possible choice of probability distribution to enforce, especially as it divides incrementing and decrementing a mode into two separate probabilities. We have chosen to do it this way for two main reasons. Firstly, modes that are least favourable to increment are most favourable to decrement, and vice versa; thus having two essentially inverse probability distributions make sense. Secondly, it allows us to sample for a fixed change in $|n|$: we select the change $\Delta|n|$, then sample for k_+ increments and k_- decrements such that $k_+ - k_- = \Delta|n|$. In this way, we can use the information from the fixed-$|n|$ algorithm to further improve our sampling. In particular, we can use the Lagrangian estimate, equation 21, as our initial starting guess, then sample away from this such that $\Delta|n|$ is weighted roughly quadratically, with a hard cutoff when we reach $|n|_{\min}$ or $|n|_{\max}$.

Algorithm 2 Algorithm for the stochastic sampling of configuration space weighted with a prior distribution defined via equation 23.

Generate some number of starting guesses in set $\mathcal{G}$.
for all Samples **do**
    Randomly select a guess $\mathbf{n} \in \mathcal{G}$
    Sample $\Delta|n|$ from $(|n| - |n|_{\text{mode}})^2$, such that $|n| \in [|n|_{\text{min}}, |n|_{\text{max}}]$.
    Randomly select k_+ up to some fixed limit, and set $k_- = k_+ - \Delta|n|$
    Sample k_+ modes to increment, k_- modes to decrement, weighted as per equation 23.
    Set $\mathbf{n}' = \hat{k}_-\hat{k}_+\mathbf{n}$.
    if $|E(\mathbf{n}') - E_{if}| \leq \delta$ **then**
        Add $\mathbf{n}'$ to $\mathcal{G}$
    end if
end for

The stochastic algorithm is described fully in Algorithm 2. We note that the $k_\pm$ modes to increment/decrement are selected *independently*, such that a mode can increase or decrease by more than one in a single sample, and it is possible to select the same sample multiple times. Algorithmically, this creates a problem of uniqueness when calculating the final value. There are two possible ways to approach this. The memory-intensive method is to set the group $\mathcal{G}$ up as a hash-table, such that a configuration $\mathbf{n}'$ only gets added to $\mathcal{G}$ if it is not already in the table. The fixed memory approach, on the other hand, is to allow $\mathbf{n}'$ to be added to $\mathcal{G}$ multiple times, then when the memory limit is reached, sort $\mathcal{G}$ and screen out duplicates. Furthermore, at this point, configurations with ρ below some threshold can either be discarded or written to file, with only the most 'important' configurations kept in $\mathcal{G}$ for the next round of sampling. The additional expense in this approach, however, is that a considerable amount of writing to and reading from file is necessary, as the final value cannot be computed until all samples have been performed.

The effectiveness of Algorithm 2 relies entirely on the fortuitous enveloping of configuration space by $\rho(\mathbf{n})$, and the increasing pool of guesses in $\mathcal{G}$. This means, however, that starting with only a single guess drastically reduces the rate at which ergodicity is approached (if ergodicity is reached at all). To this end, we use the traditional 0-1 knapsack problem fully-polynomial time approximation scheme to generate G guesses with values of

E evenly spaced between $E_{if} - \delta$ and $E_{if} + \delta$. The algorithm for doing so is well-known and described elsewhere.[14]

Finally, we must consider what number of samples will theoretically be required to achieve a sufficient characterisation of the configuration space. This is not an easy question to answer. Certainly, we would expect, from the previous section, that the relevant space will still expand exponentially even with the restrictions on $|n|$. However, we would not necessarily expect the number of configurations with significant values of ρ to increase in such a manner. Precisely we would expect the number of samples needed to follow instead the following estimate:

$$\tilde{N}_s < \int_{|n|-\Delta}^{|n|+\Delta} \rho(\mathbf{n}) \cdot \mathbf{n} \; d\mathbf{n}$$

where Δ is some measure of the spread of the distribution ρ, and we are approximating this as a continuous distribution of possible $\mathbf{n}$, hence the inequality. For the simplest possible version of the kernel in equation 13, that is

$$\rho(\mathbf{n}) = \exp\left(\sum_j c_j n_j\right) \cdot \prod_j (n_j!)^{-1} < \prod_j \frac{\exp(c_j n_j)}{n_j}$$

for some constants c_j. This equates to

$$\tilde{N}_s < \prod_j \int_0^{\bar{n}} \exp(c_j n_j) \; dn_j = \prod_j \left\{ \frac{1}{c_j} \left[\exp(c_j \bar{n}) - 1 \right] \right\}$$

where $\bar{n}$ is some estimate of the average maximum value of each n_j satisfying the range $|n| \pm \Delta$. Using the results of the previous sections, we know that $|n|_{\text{mode}} \pm \Delta \sim M/2$. This means that $|n|$ can range up to (asymptotically) MC for some constant C, making $\bar{n} \approx C$. Replacing the c_j with their average, $\bar{c}$, this becomes

$$\tilde{N}_s \sim \left(\frac{\exp(C\bar{c}) - 1}{\bar{c}} \right)^M \tag{24}$$

which is still exponential in M. However, $\bar{c}$ is typically negative; in the case of the Franck-Condon factors, $c_j = \ln y_j$, and the y_j are generally less than unity. This means that, like with the fixed-$|n|$ procedure, the mantissa is close to 1, so $\tilde{N}_s$ appears to be polynomial until M gets very large.

D. Parallelization

It is important to note that both Algorithm 1 and 2 are inherently parallelizable, and by extension therefore also the fixed-$|n|$ algorithm. Here we briefly describe the strategy for each.

The screened brute force algorithm is perhaps the most complex, because the task of determining exactly how many configurations will be screened is as complex as the enumerating itself. Certainly, we cannot simply divide the work based on the maximum n of the first mode, because the screening will necessarily result in fewer configurations for larger values of n_1. Instead, we would propose ordering the modes in terms of their maximum possible n from largest to smallest; in the case where ρ is the Franck-Condon factor, this is equivalent to in order of decreasing y_j. The simple problem of finding all bounds on the first k modes can then be solved, and the work divided up accordingly. For example, if we took the just the first two modes, for each value of n_1 we would determine the maximum n_2, and then use $n_1 n_2$ as a heuristic to assess how many configurations there will be in total. In this way a roughly equivalent number of configurations can be assigned to each thread or process, and Algorithm 1 can be carried out on each 'chunk' independently. For the fixed-$|n|$ version, the load balancing is simpler. We can quickly determine from equation 17 the maximum number of configurations for each $|n|$ from $|n|_{\min}$ to $|n|_{\max}$. As each $|n|$ can easily be performed in parallel, the task is then dividing the total number of configurations evenly between each available thread or process.

For the stochastic algorithm, every sample can in principle be done independently. However, there is the pooling of the configurations into $\mathcal{G}$ to be considered, and whether the algorithm is performed entirely in-core or with regular dumps to file. There are then two distinct possible approaches. In the first, each thread or process performs a completely independent sample run, with its own pool of guesses. The initial guess pool could be the same for each run, or be divided between all the runs. The advantage of this is that there is only overhead at the end, when all of the samples from all the threads need to be combined. This works particularly well when all valid configurations are being written to file, as the sorting and screening greatly speeds up the eventual recombination. The disadvantage is the duplicated effort, as the independent runs will likely find many of the same configurations, despite the stochasticity. The alternative then is to let all the threads share a guess pool.

Every fixed number of samples, the thread-specific configurations are broadcasted out to all the other threads, effectively pausing the sampling on all threads while the update occurs. This does not work so well with writing to file, however, as the records from each thread would need to be combined at every broadcast event, creating a considerable extra amount of file-based work.

IV. RESULTS

To test the efficiency of the new algorithms, and the validity of our asymptotic analyses, we generated a random ensemble of systems to be used with the density described in equation 2. For a selection of values of M (system size) from 20 to 50, the energies, ω_i, and weights, y_i, of each mode were selected from a uniform distribution on $[0.01, 0.41]$, and a Pareto distribution with location 10, respectively. These were chosen to resemble realistic systems, as it would be exceptionally difficult to systematically find and compute such values for an ensemble of real molecules. In particular, the Pareto distribution for the weights generally results in a few modes with large weights (greater than 0.5) while most are less than 0.1, as is found in real molecular systems. The target energy for each model system was then chosen as the mean of the ω_i multiplied by a factor drawn from a normal distribution with location 16 and unit scale. This was based on the empirical observation that for polyacenes considered elsewhere,[20,21] the excitation energy is roughly 16 times the average mode energy.

For each value of M, ten such model samples were drawn and calculated using each of the algorithms described above, with the exception that the true brute force algorithm was only performed for the lowest value of M, as it is prohibitively expensive for larger M. All the calculations were run using the open-source KNAPSACK software, running on four cores with 2 GB of memory. We also include calculations of the internal conversion rate, according to equation 5.

For the molecular systems, electronic ground state (S_0) geometries were optimized using B3LYP[27,28] and the def2-TZVP basis set,[29] in the TURBOMOLE software package.[30,31] The first singlet excited state ($S1$) was determined using TD-B3LYP with the same basis. These geometries were then used to calculate vibrational constants in the SNF package,[32] with the same functionals and basis, yielding the B-matrix in equation 3. This normal mode analysis

was combined with the forces from the earlier optimizations to generate the non-adiabatic coupling matrix elements needed in the same equation.[33] Finally, the S_0 to S_1 electronic excitation energy was calculated using DFT/MRCI[34–36] with BHLYP/def2-TZVP.[37] The reference space was iteratively generated with 10 electrons across 10 orbitals, with a maximum excitation level of 2; all electron configurations with coefficients larger than 10^{-3} were included at each step. Probe runs were calculated by discarding configurations with energy less than the highest reference energy; starting with a barrier of 0.6 E_h, then 0.8, with finalised wavefunction built using a barrier of 1.0 E_h. Molecular orbitals with energies larger than 2.0 E_h were not used. The number of modes included in the rate calculation was reduced by two means. Firstly, a number of modes are equivalent by point group symmetry, typically reducing the number of modes by a factor of 2 (the order of the point group). Secondly, any modes with a Huang-Rhys factor of less than 10^{-6}.

A. Model systems

For every model system, the brute force (where applicable), hybrid brute force, and fixed-$|n|$ algorithms all gave values of the total density $\sum_{\mathbf{n}} \rho(\mathbf{n})$ that agreed to within a thousandth of a percent. For the stochastic algorithm, we first need to ascertain the appropriate number of samples to achieve convergence to the correct result. This was done by systematically increasing the number of samples from 10^8 for a single model for each value of M, and determining at what number the density plateaus. Figure 4 shows this analysis for the three smallest system sizes.

From the figure, we see that the density does flatten out, and approaches the correct value ascertained from the non-stochastic algorithms. However, the deviation of the value after 'convergence' is typically on the order of 0.1%, and adding additional samples does not help. The reason for this is sensitivity to the pool of starting guesses - if the total $|n|$ for each starting guess is too large, the algorithm 'walks' to higher and higher $|n|$, missing the most important configurations as would be located using the fixed-$|n|$ algorithm. As such, when generating the starting guesses as per step 1 of Algorithm 2, we took the following approach. The energy window was divided into 20 equally spaced chunks. We then used the fully polynomial time algorithm for solving the 0-1 knapsack problem to find configurations close to each energy division. From this best guess, we scan for the nearest acceptable

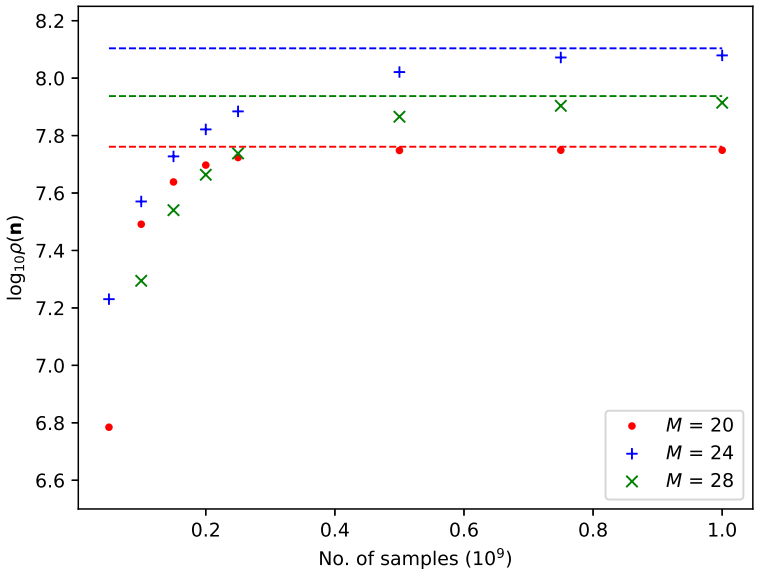

FIG. 4. Convergence of the value of the density, ρ, with the number of samples used in the stochastic algorithm, for a selection of the model systems described in the main text. The extrapolated asymptotes for each system are shown as dashed lines.

configuration with total $|n|$ less than or equal to the target modal value of $|n|$, as determined by the Lagrangian method. This scan can be achieved rapidly by energy ordering the modes and iteratively adding or subtracting occupations so as to not change the energy by more than a given tolerance, while approaching the correct total $|n|$. In practice we have seen that this can typically be found in no more than 20 iterations. Seeding the guess pool in this way led to the promising results of Figure 4; in contrast, using fewer guesses or not adjusting the value of $|n|$ appropriately, led to considerable underestimates of the value of the density, sometimes of several orders of magnitude.

Having determined the appropriate number of samples to use in the stochastic algorithm, we can now look at the scaling of both the computational time and number of significant configurations found using each algorithm. We show these in Figure 5. Note that we expect the hybrid brute force to find the most configurations, with consistently more configurations screened out due to insignificance in both the fixed-$|n|$ and stochastic algorithms. The amount of computational time should naturally scale roughly linearly with the number of configurations, however the stochastic algorithm has a considerable overhead due to needing

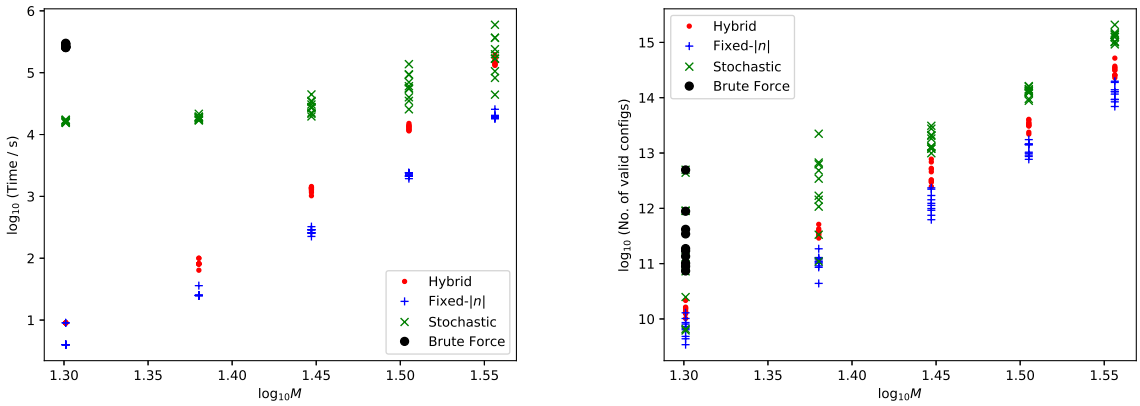

FIG. 5. Log-log plots of the scaling in computational wall time (left) and number of significant configurations (right) for the ensemble of model systems described in the main text. Full brute force results (black circles) are given only for the lowest value of M, as based on the number of configurations that would need to be checked for the next biggest, such a calculation would take several millennia.

to sort and write to file configurations whenever the allocated memory is full. This is due to the redundancy inherent in the algorithm.

We see these trends clearly in Figure 5. Importantly, both the hybrid (screening-only) and fixed-$|n|$ deterministic algorithms appear to show strictly linear log-log scaling of number of valid configurations, which implies polynomial behaviour with respect to increasing system size. This demonstrates that the density of significant configurations (that also satisfy the energy criterion) is not in fact exponential. The time taken - which encompasses all configurations checked, not just those which contribute significantly - is not linear for the hybrid algorithm, however, with a clearly discernible slight increase in gradient with increasing M, indicating exponential behaviour. The fixed-$|n|$ algorithm, on the other hand, appears to be linear over the range of M considered, with a straight line fit (with gradient of 14.4) explaining over 99% of the observed variance. This reflects the predicted quasi-polynomial behaviour due to restricting the configuration space.

The results for the stochastic algorithm are less clear. There is a much greater spread of timings within each set of model systems, due to the number of samples being required to reach convergence differing between them. On the other hand, the number of significant configurations stays fairly consistent for each value of M, suggesting that the variability

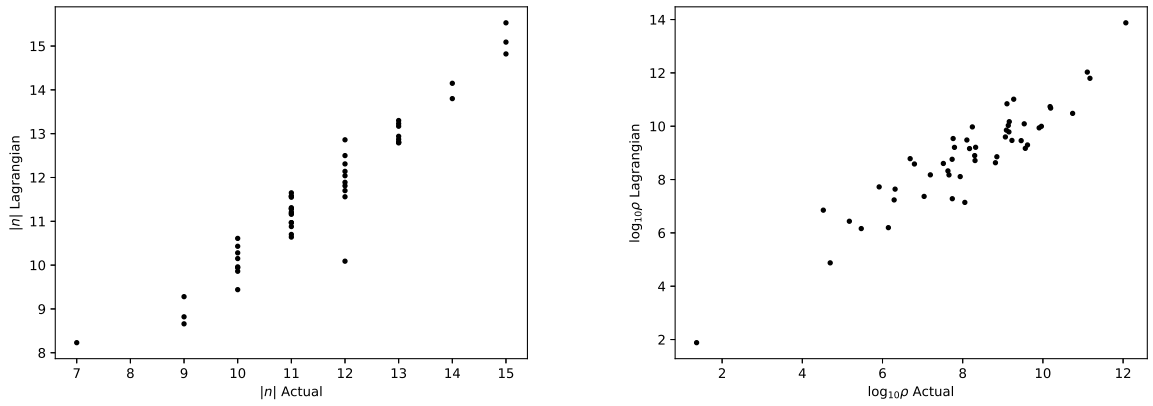

FIG. 6. Comparisons of the modal value of $|n|$ (left) and total density ρ (right)as predicted by the Lagrangian procedure, versus that found by tabulating all configurations from the fixed-$|n|$ algorithm.

is largely due to the ease of finding suitable starting guesses. Interestingly, the number of configurations is higher for the stochastic algorithm than all but the full brute force algorithm, which implies that our importance criterion is not strong enough, and a more efficient sampling could be achieved by tightening this. However, this effect lessens with increasing M, and most importantly there is clearly a point at which the time taken for the stochastic algorithm will be considerably less than for the other algorithms. In fact, this has already started to happen for the largest value of M considered in Figure 5. This is because the cost of each individual sample does not increase with system size (beyond a small amount of additional overhead in the sorting steps), whereas the recursive nature of the other algorithms mean they get far more memory-intensive and thus expensive as M increases. Finally, we note that all three new methods perform orders of magnitude better than the full brute force approach for the smallest system (10^5 times faster in the case of the fixed-$|n|$ algorithm), and this improvement will only increase as M does, given the factorial scaling of the naive method.

It is also worth looking at the Lagrangian approximation of equation 21 in more detail, as it affects the efficacy of both the fixed-$|n|$ and stochastic algorithms. Figure 6 shows the predicted modal $|n|$ and the estimated density ρ as determined using this approach, compared to that found from the hybrid brute force algorithm, across all model systems. Generally speaking, the Lagrangian prediction of $|n|$ is within 1 of the empirical mode, and

is thus a very effective prediction for the overall distribution of significant $|n|$. However, the approximation to the density is very poor; it is on average two orders of magnitude too large or too small, with the nature of the error depending on the exact distribution of weights and eneriges. As such it neither provides a convenient bound nor good approximation to the density. This is important in the context of non-radiative rates as this method has been used previously to estimate rates,[22] which being directly proportional to this density, will also likely be in error by orders of magnitude.

B. Example rates

As an important final test, we apply the new algorithms to the determination of internal conversion rates for three molecular systems: anthracene, tetracene, and indole. We have selected these because they are molecules that are known to have a viable internal conversion pathway,[38–41] and that have a feasible number of vibrational modes for the calculations. Our intention here is not to assess the validity of the physical assumptions underlying the formalism inherent in equation 5, as that is not the topic of this paper, and has been considered elsewhere.[20,42] In fact, it is very difficult to validate specifically the calculated internal conversion rate for a single pathway, as there is very limited experimental data for molecules that we can treat computationally, and moreover these experiments will typically be in solvent. Instead, we are demonstrating that for realistic systems, where the relevant parameters have not been randomly estimated as per the model ensembles above, that the new algorithms give consistent results.

In Table I, we tabulate the results from these algorithmic comparisons, along with a summary of the relevant parameters. Crucially, we see that the hybrid and fixed-$|n|$ algorithms consistently give the same results, with values of $|n|_{\max}$ and $|n|_{\mathrm{mode}}$ determined more by the target transition energy than by the dimension M. As it is this parameter that controls the computational cost of the fixed-$|n|$ method, this is promising. However, the number of configurations with a Franck-Condon factor greater than the fixed threshold (here chosen as 10^{-12}) does increase noticeably with M. This is a particular problem for the stochastic method, where we see that in the case of indole, convergence has not quite been achieved even at 2.5×10^{10} samples. However, the stochastic results are consistently much closer to the deterministic results than those found through the Lagrangian approximation. This is

TABLE I. Internal conversion rates as calculated using equation 5 for anthracene, tetracene, and indole, for each of the algorithms described in this work. Additionally, we give the number of modes M, target energy E, and mode-weight parameter α for each, along with algorithm parameters. The energy window δ in all cases was 40 meV. The brute force rate for anthracene was $2.120 \times 10^5\ s^{-1}$.

	Anthracene	Tetracene	Indole
M	12	22	29
$E_{S_0 \to S_1}$ (eV)	3.172	2.435	4.689
α	1.8	2.5	2.4
$\|n\|_{\min}$	11	8	10
$\|n\|_{\max}$	18	16	24
$\|n\|_{\text{mode}}$	14	12	19
No. samples	1×10^8	5×10^9	2.5×10^{10}
No. config.	2.89×10^5	2.02×10^8	3.07×10^9
		$k_{IC,FC}\ (s^{-1})$	
Hybrid	2.120×10^5	5.276×10^6	1.258×10^9
Fixed-$\|n\|$	2.120×10^5	5.276×10^6	1.258×10^9
Stochastic	2.094×10^5	5.080×10^6	9.716×10^8
Lagrangian	1.825×10^5	1.142×10^7	6.885×10^3

particularly severe in the case of Indole where the sheer number of significant configurations has resulted in the Lagrangian result being six orders of magnitude too small. Finally, we see that the fractional energy parameter, α, which controls the scaling as per equation 18, is close to but less than e. This effectively means that, as expected, the scaling is still strictly exponential but for this size of systems, the cost is not yet prohibitive. These results are therefore promising for the application of these new algorithms to much larger systems.

V. CONCLUSIONS

In this work, we have developed three novel methods for the enumeration of bosonic configurations to calculate a density kernel with generalized Boltzmannian statistics. The

two deterministic algorithms, described by Algorithm 1, are based, respectively, on heavy screening of the configurations and division of the simplicial volume into fixed occupation slices. The third, stochastic approach, given in Algorithm 2, is based on an importance sampling generated from a pool of guesses, using heuristic solutions to the knapsack problem as a starting point. Our focus in validation has been on applying these new methods to the problem of calculating non-radiative rate constants for molecules, but the algorithms have been designed to be agnostic to the specific choice of physical problem. Our asymptotic analysis, and subsequent numerical investigations, demonstrate that the seemingly factorial scaling of these combinatorial problems is unphysical. In particular, for an ensemble of model systems with the simplest Boltzmann kernel, the number of significant configurations in fact seems to scale as a high-order polynomial. This has important consequences for future algorithmic developments, as it suggests that by improving the choice of importance sampling procedure, the problem can eventually be solved in polynomial time.

The efficient sampling and achievement of ergodicity in the stochastic approach is not simple, however, and there is still clearly much that could be improved. In particular, the sheer volume of the space to be sampled is so staggeringly large that the choice of appropriate seeding guesses is itself a computationally intensive challenge. This is alleviated somewhat by our observations of where the maximum total occupation can be found, as reflected in the fixed-$|n|$ algorithm, as this necessarily reduces the space to be searched. It is not sufficient, however, to limit ourselves only to the distributional maximum, as demonstrated by the poor performance of the Lagrangian approximation in many cases. One possible approach would be to solve a simpler problem in the subspace of k modes with the largest weights, and then perturbatively expand into the full space from there. This would then allow for more efficient directing of the sampling procedure to the important regions of configuration space. The difficulty comes from defining where the cutoff for the size of the subspace should be, and assessing how this affects the final results.

Similarly, the relevant molecules in new materials design are typically much larger than the three molecular systems we have considered here, with M typically being in the range of 50 to 100, after symmetry and insignificant Huang-Rhys factors have been taken into account. We did not include such large molecules because of the sheer amount of computational effort in obtaining the necessary parameters to a sufficient accuracy, although we hope to do this in a follow-up study. Our model tests do show scalability up to the lower end of this range of M,

and we would anticipate from extrapolating the scaling behaviours that it should be feasible to consider these larger systems. However, it is an open question whether the sampling will be sufficient for the highest values of M. It would also be worthwhile to compare how changing the choice of polynomials in the kernel definition, equation 13, affects the scaling parameter, α, and whether this can be used to improve sampling efficiency. These are mathematically challenging questions that suggest interesting pathways for future developments.

SUPPLEMENTARY MATERIAL AND DATA AVAILABILITY

Open-source implementations of all the methods described here can be found in the Knapsack software in the following GitHub repository: `https://www.github.com/robashaw/knapsack` (Last accessed: 19th November 2020) The input parameters for all of the molecular and model systems can be found in the supplementary material CSV file, along with XYZ coordinates for the ground-state molecular systems. Additional derivations of results can be found in the supplementary material PDF.

ACKNOWLEDGMENTS

This work was supported by the Australian Government through the Australian Research Council (ARC) under the Centre of Excellence scheme (project number CE170100026). It was also supported by computational resources provided by the Australian Government through the National Computational Infrastructure National Facility and the Pawsey Supercomputer Centre.

REFERENCES

[1]J. P. Sethna, *Statistical mechanics: entropy, order parameters, and complexity*, Oxford master series in statistical, computational, and theoretical physics No. 14 (Oxford University Press, Oxford ; New York, 2006) oCLC: ocm63136230.

[2]B. Hirshberg, V. Rizzi, and M. Parrinello, "Path integral molecular dynamics for bosons," Proceedings of the National Academy of Sciences **116**, 21445–21449 (2019).

[3]J. Katriel, R. Pauncz, and J. J. C. Mulder, "Studies in the configuration interaction method. II. Generating functions and recurrence relations for the number of many-particle configurations," International Journal of Quantum Chemistry **23**, 1855–1867 (1983).

[4]A. Torre, L. Lain, and J. Millan, "Calculation of the dimension of full configuration interaction spaces: application to the determination of spectroscopic terms," Journal of Molecular Structure: THEOCHEM **287**, 63–66 (1993).

[5]A. I. Streltsov, O. E. Alon, and L. S. Cederbaum, "General mapping for bosonic and fermionic operators in Fock space," Physical Review A **81**, 022124 (2010).

[6]J.-C. Pain and M. Poirier, "Analytical and numerical expressions for the number of atomic configurations contained in a supershell," Journal of Physics B: Atomic, Molecular and Optical Physics **53**, 115002 (2020).

[7]D. F. Heller, K. F. Freed, and W. M. Gelbart, "Dependence of Radiationless Decay Rates in Polyatomic Molecules upon the Initially Selected Vibronic State: General Theory and Application," The Journal of Chemical Physics **56**, 2309–2328 (1972).

[8]D. F. Heller and K. F. Freed, "Energy dependence of nonradiative decay in polyatomic molecules," International Journal of Quantum Chemistry **6**, 267–277 (1972).

[9]V. G. Plotnikov, "Regularities of the processes of radiationless conversion in polyatomic molecules," International Journal of Quantum Chemistry **16**, 527–541 (1979).

[10]M. Bixon and J. Jortner, "Intramolecular Radiationless Transitions," The Journal of Chemical Physics **48**, 715–726 (1968).

[11]S. Martello and P. Toth, *Knapsack problems: algorithms and computer implementations*, Wiley-Interscience series in discrete mathematics and optimization (J. Wiley & Sons, Chichester ; New York, 1990).

[12]J. Franck and E. G. Dymond, "Elementary processes of photochemical reactions," Transactions of the Faraday Society **21**, 536 (1926).

[13]H. Kellerer, U. Pferschy, and D. Pisinger, "Introduction to NP-Completeness of Knapsack Problems," in *Knapsack Problems* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004) pp. 483–493.

[14]V. V. Vazirani, *Approximation algorithms*, corrected second printing ed. (Springer, Berlin Heidelberg, 2003) oCLC: 249006472.

[15]K. E. Knowles, E. A. McArthur, and E. A. Weiss, "A Multi-Timescale Map of Radiative and Nonradiative Decay Pathways for Excitons in CdSe Quantum Dots," ACS Nano **5**,

2026–2035 (2011).

[16]C. Hauswald, P. Corfdir, J. K. Zettler, V. M. Kaganer, K. K. Sabelfeld, S. Fernández-Garrido, T. Flissikowski, V. Consonni, T. Gotschke, H. T. Grahn, L. Geelhaar, and O. Brandt, "Origin of the nonradiative decay of bound excitons in GaN nanowires," Physical Review B **90**, 165304 (2014).

[17]D. Noy, C. C. Moser, and P. L. Dutton, "Design and engineering of photosynthetic light-harvesting and electron transfer using length, time, and energy scales," Biochimica et Biophysica Acta (BBA) - Bioenergetics **1757**, 90–105 (2006).

[18]K. Wang, Z. Gao, W. Zhang, Y. Yan, H. Song, X. Lin, Z. Zhou, H. Meng, A. Xia, J. Yao, and Y. S. Zhao, "Exciton funneling in light-harvesting organic semiconductor microcrystals for wavelength-tunable lasers," Science Advances **5**, eaaw2953 (2019).

[19]F. Häse, L. M. Roch, P. Friederich, and A. Aspuru-Guzik, "Designing and understanding light-harvesting devices with machine learning," Nature Communications **11**, 4587 (2020).

[20]R. R. Valiev, V. N. Cherepanov, G. V. Baryshnikov, and D. Sundholm, "First-principles method for calculating the rate constants of internal-conversion and intersystem-crossing transitions," Physical Chemistry Chemical Physics **20**, 6121–6133 (2018).

[21]R. R. Valiev, V. N. Cherepanov, G. V. Baryshnikov, and D. Sundholm, "Calculating rate constants for intersystem crossing and internal conversion in the Franck–Condon and Herzberg–Teller approximations," Physical Chemistry Chemical Physics **21**, 6121–6133 (2018).

[22]R. R. Valiev, R. T. Nasibullin, V. N. Cherepanov, G. V. Baryshnikov, D. Sundholm, H. Ågren, B. F. Minaev, and T. Kurtén, "First-principles calculations of anharmonic and deuteration effects on the photophysical properties of polyacenes and porphyrinoids," Physical Chemistry Chemical Physics **22**, 22314–22323 (2020).

[23]R. A. Shaw, "Knapsack v1.0," (2020), https://www.github.com/robashaw/knapsack, Last Accessed 3rd December 2020.

[24]S. Banerjee, A. Baiardi, J. Bloino, and V. Barone, "Temperature Dependence of Radiative and Nonradiative Rates from Time-Dependent Correlation Function Methods," Journal of Chemical Theory and Computation **12**, 774–786 (2016).

[25]H. S. M. Coxeter, *Regular polytopes*, 3rd ed. (Dover Publications, New York, 1973).

[26]D. A. McQuarrie, *Statistical mechanics* (University Science Books, Sausalito, Calif, 2000).

[27]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," Physical Review A **38**, 3098–3100 (1988).

[28]C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," Physical Review B **37**, 785–789 (1988).

[29]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005), publisher: The Royal Society of Chemistry.

[30]"TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from `https://www.turbomole.org`.".

[31]S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. Marefat Khah, S. Karbalaei Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński, and J. M. Yu, "TURBOMOLE: Modular program suite for *ab initio* quantum-chemical and condensed-matter simulations," The Journal of Chemical Physics **152**, 184107 (2020).

[32]J. Neugebauer, M. Reiher, C. Kind, and B. A. Hess, "Quantum chemical calculation of vibrational spectra of large molecules?Raman and IR spectra for Buckminsterfullerene," Journal of Computational Chemistry **23**, 895–910 (2002).

[33]R. Send and F. Furche, "First-order nonadiabatic couplings from time-dependent hybrid density functional response theory: Consistent formalism, implementation, and performance," The Journal of Chemical Physics **132**, 044107 (2010).

[34]S. Grimme and M. Waletzke, "A combination of Kohn–Sham density functional theory and multi-reference configuration interaction methods," The Journal of Chemical Physics **111**, 5645–5655 (1999).

[35]I. Lyskov, M. Kleinschmidt, and C. M. Marian, "Redesign of the DFT/MRCI Hamiltonian," The Journal of Chemical Physics **144**, 034104 (2016).

[36]A. Heil and C. M. Marian, "DFT/MRCI Hamiltonian for odd and even numbers of electrons," The Journal of Chemical Physics **147**, 194104 (2017).

[37]A. D. Becke, "A new mixing of Hartree–Fock and local density-functional theories," The Journal of Chemical Physics **98**, 1372–1377 (1993).

[38]M. Baba, M. Saitoh, K. Taguma, K. Shinohara, K. Yoshida, Y. Semba, S. Kasahara, N. Nakayama, H. Goto, T. Ishimoto, and U. Nagashima, "Structure and excited-state dynamics of anthracene: Ultrahigh-resolution spectroscopy and theoretical calculation," The Journal of Chemical Physics **130**, 134315 (2009).

[39]A. Kearvell and F. Wilkinson, "Internal conversion from the lowest excited singlet states of aromatic hydrocarbons," Chemical Physics Letters **11**, 472–473 (1971).

[40]Y. Harabuchi, K. Saita, and S. Maeda, "Exploring radiative and nonradiative decay paths in indole, isoindole, quinoline, and isoquinoline," Photochemical & Photobiological Sciences **17**, 315–322 (2018).

[41]G. Díaz Mirón and M. C. González Lebrero, "Fluorescence Quantum Yields in Complex Environments from QM-MM TDDFT Simulations: The Case of Indole in Different Solvents," The Journal of Physical Chemistry A **124**, 9503–9512 (2020).

[42]P. V. Komarov and V. G. Plotnikov, "Influence of intermolecular interactions on spectral-luminescent properties of a polyatomic molecule," International Journal of Quantum Chemistry **112**, 3039–3045 (2012).