Algorithm for efficient elastic transport calculations for arbitrary device geometries

Douglas J. Mason, David Prendergast, Jeffrey B. Neaton, and Eric J. Heller

Phys. Rev. B **84**, 155401 — Published 3 October 2011

DOI: 10.1103/PhysRevB.84.155401

Efficient Elastic Transport Calculations for Arbitrary Device Geometries

Douglas J. Mason
Department of Physics, Harvard University, Cambridge, MA 02138, USA and
Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

David Prendergast and Jeffrey B. Neaton
Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

Eric J. Heller
Department of Physics, Harvard University, Cambridge, MA 02138, USA

With the growth in interest in graphene, controlled nanoscale device geometries with complex form factors are now being studied and characterized. There is a growing need to understand new techniques to handle efficient electronic transport calculations for these systems. We present an algorithm that dramatically reduces the computational time required to find the local density of states and transmission matrix for open systems regardless of their topology or boundary conditions. We argue that the algorithm, which generalizes the recursive Green's Function method by incorporating the reverse Cuthill-McKee algorithm for connected graphs, is ideal for calculating transmission through devices with multiple leads of unknown orientation, and becomes a computational necessity when the input and output leads overlap in real space. This last scenario takes the Landauer-Buttiker formalism to general scattering theory in a computational framework that makes it tractable to perform full-spectrum calculations of the quantum scattering matrix in mesoscopic systems. We demonstrate the efficacy of these approaches on graphene stadiums, a system of recent scientific interest, and contribute to a physical understanding of Fano resonances which appear in these systems.

PACS numbers:

I. INTRODUCTION

In this work we describe a method for computing the electron transport properties of systems with arbitrary geometries, focusing on single layers of carbon known as graphene. Recent experiments have stepped outside the realm of linear, rectangular MOSFET-type devices and into fully two-dimensional geometries with multiple leads at arbitrary angles[11,19,25]. Challenges to graphene fabrication give even linear devices substantial two-dimensional character, largely due to unpredictable defects in the etching process[20,23]. Moreover, novel applications for graphene have been recently proposed using spin polarization that explicitly relies on irregular geometries[27]. The current gap between theory and experiment in the literature can be attributed to the lack of efficient computational tools to handle such arbitrary devices at the nano- and meso-scopic scales. This paper aims to provide one such tool.

We provide an algorithm that generalizes the well-known recursive Green's function (RGF) method outlined by Datta[5] by incorporating the reverse Cuthill-McKee algorithm for connected graphs[3]. By reinterpreting the Landauer-Buttiker formalism, we demonstrate that RGF can work for systems that do not fit an input-system-output schematic, expanding the fruits of algorithmic advances in transmission calculations to the general scattering matrix problem. Like RGF, our algorithm can also produce the local density of states (LDOS) at a comparable computational cost.

One important aim for this paper is to produce and explain a method that is straightforward to implement and does not require cumbersome external software packages. Accordingly, each calculation is performed using the standard BLAS and LAPACK dense matrix algebra routines[7] which come pre-installed on nearly all scientific machines today. Comparisons to sparse matrix packages like SuperLU[8] are discussed in Section IV D.

II. BACKGROUND

The methods used in this article apply to Hamiltonians with sparse character (composed of many off-diagonal zeros) and poses some order of localization, that is, one part of the Hamiltonian doesn't couple to another distant part. For instance, for all of our calculations we will be using the general single-orbital nearest-neighbor tight-binding Hamiltonian

$$H = \sum_i u_i \mathbf{a}_i^\dagger \mathbf{a}_i + \sum_{\{i,j\}} t_{ij} \mathbf{a}_i^\dagger \mathbf{a}_j \qquad (1)$$

where $\mathbf{a}_i$ is the annihilation operator for the i^{th} orbital and the set $\{i,j\}$ cycles through all nearest-neighbor pairs. In graphene, $u_i = 0$ and $t_{ij} = 2.7$eV which sets the value of the Fermi level at $E = 0$eV. These methods are quite general since the structure of this Hamiltonian is very similar to any finite-difference continuous wave equation sampled on a lattice of any character. Moreover, recursive methods apply to all other orders of tight-binding and finite-difference approximations, such as nearest-nearest-neighbor tight-binding, although their efficiency will drop for increasing orders. It has also been shown that such methods can apply to interacting Hamiltonians under certain conditions[17].

When calculating quantum transport across a device, the Hamiltonian is considered to be infinite. In the two-terminal case, a left lead is described by the Hamiltonian H_L, a right lead by H_R, and a central region by H_C. We have

$$H = \begin{pmatrix} H_L & V_{LC} & 0 \\ V_{CL} & H_C & V_{CR} \\ 0 & V_{RC} & H_R \end{pmatrix} \tag{2}$$

To render the problem tractable, methods have been devised to collapse the system onto a finite Hamiltonian of the same dimension as the central region by projecting the contribution from the leads onto the central region. For a lead modeled by an infinitely-repeated unit cell, we use the algorithm of[13,14] which provides state-of-the-art efficiency. Each lead is thus represented by an energy-dependent retarded self-energy $\Sigma_i^r(E)$, which has the same dimension as the central region. For the Hamiltonians of interest to this paper, the retarded self-energy for each lead is a zero-matrix except at the boundary on the central region where it meets the lead. We represent the contribution from all leads as the sum $\Sigma^r(E) = \sum_i \Sigma_i^r(E)$ as in[15]. This gives us a Hamiltonian describing the infinite system as a finite Hamiltonian

$$H'(E) = H_C + \Sigma^r(E)$$

which has the same dimension as the central (finite) region. Because H_C describes the central region as a closed system, it is Hermitian. Accordingly, H' is Hermitian except where there are contributions from the self-energies of the leads.

A. Calculating the LDOS and Transmission Matrix from the Green's Function

The LDOS and transmission matrix are calculated from the non-interacting Green's function matrix defined by

$$G(E) = \left[(E + i\eta) \cdot \mathbb{I} - H\right]^{-1} \tag{3}$$

Here $E + i\eta$ is the energy of the electron being scattered, subtracting a small imaginary parameter designed to avoid poles in the complex plane. Because of the imaginary contribution from the self-energy of the leads, we can omit this small quantity and use

$$G(E) = \left[E \cdot \mathbb{I}_C - H'(E)\right]^{-1} \tag{4}$$

where the quantity $\mathbb{I}_C$ is the identity matrix with the dimension of the central region. In the tight-binding approximation, the diagonal entries of H' will give the on-site energy of each atomic orbital and the off-diagonal elements will give the hopping potential between neighboring orbitals. Using this formalism, we can identify the local density of states as

$$D(E, n) = \frac{1}{\pi} \text{Im}\left[G_{n,n}(E)\right] \tag{5}$$

That is, the local density of states is encoded along the diagonal entries of the full retarded Green's function of the system. From here on out, we omit the explicit energy dependence in G.

The transmission matrix is calculated not from entries of G along the diagonal but from the off-diagonal elements communicating information from the input to the output boundaries. Even though these boundaries can be general, we choose to use the familiar two-terminal nomenclature, so that we write the relevant sub-matrix as G_{LR}. Similarly, we write the self-energies for the input and output boundaries as $\Sigma^r_{L,R} = \sum_{i \in L,R} \Sigma^r_i$. We assume that we have obtained the self-energies for all boundaries, and that our incoming and outgoing wavefunctions are the open modes of those boundaries. Accordingly, we define these modes as $\Gamma_{L,R} = 2\text{Im}\left[\Sigma^r_{L,R}\right]$. The incoming and outgoing wavefunctions for each mode can then be represented by the matrix square-root of the gamma matrix, and the transmission function between these modes can be written as a matrix

$$t = \left(\Gamma_\text{L}\right)^{1/2} G_\text{LR} \left(\Gamma_R\right)^{1/2} \tag{6}$$

In the linear regime, conductance through the system will be proportional to the sum-squares of transmission functions for each incoming mode. Trace identities then produce the transmission probability[17]

$$T(E) \propto \text{Tr}\left[T\right] = \text{Tr}[tt^\dagger] = \text{Tr}\left[\Gamma_L G^*_{LR} \Gamma_R G_{LR}\right] \tag{7}$$

B. Full Inversion

The most straightforward calculation of the LDOS and transmission matrix involves first inverting the entire matrix $M = E \cdot \mathbb{I}_C - H'$ and then projecting out the diagonal elements of G for the LDOS and the sub-matrix G_{LR} for the transmission matrix. As is well known, the time to compute the inverse of a matrix with N rows and columns scales as N^3. Moreover, the sparse systems this paper addresses invert to dense matrices, adding a memory cost that scales by N^2. Such large scaling factors make this calculation prohibitively costly for systems on the order of thousands of atoms.

A shortcut to calculating the LDOS and transmission can be made by solving a set of linear equations $M\vec{x}_i = \hat{e}_{i,C}$ where $\hat{e}_{i,C}$ is a unit vector of size C with unity on the basis index i. Solving for the diagonal entries then requires solving for $\vec{x}_i$ for all $i \in [1, N]$. Such calculations can be aided by sophisticated sparse matrix software packages which cut the number of operations by permuting the matrix columns and rows, and many different approaches are outlined in[7]. Solving for all the diagonal entries, however, requires one to solve for N separate systems of equations, which makes these approaches less efficient than one would hope. For instance it can be shown that nested-dissection methods[6] can under optimal circumstances return the inverse with scaling of order $N^2 \log N$ after a re-ordering operation whose cost grows with some function of N. This is better than N^3 but still worse than N^2 as promised by the linear recursive Green's Function method.

C. Linear Recursive Green's Function Method

To reduce the computational footprint of LDOS and transmission calculations, the recursive Green's Function method was developed. In its usual implementation, the recursive Green's Function method operates on a Hamiltonian that satisfies the following three conditions:

1. An input lead contributes the boundary condition Σ_L^r from the left (incoming)

2. An output lead contributes the boundary condition Σ_R^r from the right (outgoing)

3. A linear device rests in between the leads, which can be divided into N vertical slices referred to as "primary layers", which we number in increasing order from left-to-right.

While the particular expression of this topology can be distorted, the means of calculation is always the same, and assumes that the system can be mapped onto a linear chain of primary layers. Accordingly, we refer to it as the linear recursive Green's Function method, or LRGF. In LRGF, one employs the Dyson equation, which is derived from partial block inversion, to move left-to-right along the primary layers (see, for example,[5,24]). In fact, LRGF performs partial block inversion on the tridiagonal matrix

$$M = \begin{pmatrix} H'_1 & H'_{12} & & & \\ H'_{21} & H'_2 & H'_{23} & & \\ & H'_{32} & \ddots & \ddots & \\ & & \ddots & H'_{N_L-1} & H'_{N_L-1,N_L} \\ & & & H'_{N_L,N_L-1} & H'_{N_L} \end{pmatrix} \quad (8)$$

where H'_i is the system Hamiltonian H' projected at the primary layer i. The partial-block inversion algorithm is outlined in the next section.

Algorithm 1 Diagonal-block-inversion for block-tridiagonal matrix

```
1. $g_1^L = M_1^{-1}$
2. for $i = 2$ to $N_L$
    (a) $\Sigma_i^L = M_{i,i-1} g_{i-1}^L M_{i-1,i}$
    (b) $g_i^L = \left(M_i - \Sigma_i^L\right)^{-1}$
3. end for
4. $G_{N_L} = g_{N_L}^L$
5. $g_{N_L}^R = M_{N_L}^{-1}$
6. for $i = N_L - 1$ to 1
    (a) $\Sigma_i^R = M_{i+1,i} g_{i+1}^R M_{i,i+1}$
    (b) $g_i^R = \left(M_i - \Sigma_i^R\right)^{-1}$
    (c) $G_i = g_i^L \left(\mathbb{I}_i - \Sigma_i^R g_i^L\right)^{-1}$
7. end for
```

D. Block Inversion

To demonstrate the block inversion algorithm, we rewrite

$$M = \begin{pmatrix} M_1 & M_{12} & & \\ M_{21} & M_2 & \ddots & \\ & \ddots & \ddots & M_{N_L-1,N_L} \\ & & M_{N_L-1,N_L} & M_{N_L} \end{pmatrix} \quad (9)$$

and employ Algorithm 1.

The first for-loop (Step 2) returns the inverse $G_{N_L} = \left(M^{-1}\right)_{N_L}$ at the bottom-right-most block (see Step 4). The second for-loop (Step 6) returns the set of block inverses along the diagonal: $G_i = \left(M^{-1}\right)_i$ of sizes $\{N_i | \sum_i N_i = N\}$. From these blocks one can obtain the diagonal elements of M^{-1} and thus the LDOS. The efficiency of this algorithm scales $\leq \max_i N_i^3 N_L$ and in the case of a square device, where $\{N_i\} \sim N_L \sim \sqrt{N}$, it can scale with N^2, a vast improvement over full inversion when only the diagonal entries are required. A further extension allows us to calculate any block of the inverse off the diagonal, but at the cost of additional operations (see Cauley et al.[2]). For LRGF, calculating G_{LR} requires us to calculate the inverse at the far upper-right block $\left(M^{-1}\right)_{1,N_L}$ and an additional step in the first for-loop can compute this block while minimizing memory allocation (see Datta[5]).

III. THE OUTWARD WAVE METHOD

For linear systems satisfying the conditions of LRGF (see Section II C) the block-tridiagonal nature of $M = E \cdot \mathbb{I}_C - H'$ is evident: one simply slices the device into vertical sections, from left-to-right. However, this is not the case for general geometries even though many Hamiltonians are sparse and exhibit a similar structure. In order to take advantage of the computational efficiency of LRGF, one must find a way to map the system geometry onto a linear chain,

Literature through the past two decades describes many inventive methods to accomplish precisely this goal. Among those methods are a conformal map to transform a quasi-circular system onto a linear chain, using continuous eigenfunctions as their basis[26], and a unique geometry for applying LRGF to four-terminal devices[1]. More recent work on the contact-block reduction method[16] divides a generic device into smaller blocks which are pieced together like a jig-saw puzzle. In addition, graph theory has been used to develop a relatively elaborate system permitting the use of LRGF with generic boundary conditions[28]. These results, along with others[21], suggest the approach we explore in depth in this article, however we argue that the formalism of the "virtual lead" is not necessary. Our formalism, in addition, opens the Landauer-Buttiker formalism to tractable reflection matrix calculations as described in Section III E.

A. Reverse Cuthill-McKee

Given any sparse matrix A, the Reverse Cuthill-McKee (RCM) algorithm[4] automatically calculates a permutation matrix P so that PAP^T produces a block-tridiagonal matrix, which enables us to use the LRGF method. The only requirement for RCM is that the matrix A satisfy the properties of an adjacency matrix, which describes the edges between vertices of an undirected graph. This is satisfied when the non-zero entries of a matrix are symmetrically distributed across the diagonal, and is therefore satisfied for any Hermitian Hamiltonian. Since the tight-binding and similar localized models create Hamiltonians that describe actual graphs, where nodes map onto atomic orbitals and edges onto overlap functions which are distributed in physical space, RCM is ideal for such systems.

RCM aims to minimize the distance of non-zero entries to the diagonal, which makes it a "bandwidth minimization" algorithm and ideal for our purposes. This is because the most computationally expensive step in an LRGF calculation constitutes inverting each individual block, and this time is dependent upon the cube of number of rows N_i for each block. However, we are constrained by the fact that the number of rows for each block must add up to the number of rows in the entire system, that is, $\sum_i N_i = N$. We can write a rough optimization function $\sum_i N_i^3$ which describes the time of calculation. This optimization function is minimized when the number of blocks is maximized, and the size of each block is reduced. Ideally, no block is especially large compared to the others, since the cubic function grows rapidly.

Algorithm 2 Reverse Cuthill-McKee

```
1. define $S_1, i = 2$
2. while $S_{i-1} \neq \emptyset$
    (a) define $S_i$ as the indices of the columns of the off-diagonal elements in the rows $S_{i-1}$
    (b) $S_i = S_i / \{S_j | j = 1, \ldots, i-1\}$ that is, eliminate the indices that have been visited previously
    (c) $i = i + 1$
3. end while
4. reverse subscripts of $\{S_i\}$
```

Figure 1: The six sites of a model nearest-neighbor tight-binding system are shown with index labels. Lines between sites indicate off-diagonal entries in the Hamiltonian

RCM is able to reduce block sizes by keeping track of site indices while propagating through the system like a wave propagates on a pond surface. It begins by taking a seed of indices S_1, which constitute a set of nodes in the graph represented by A. RCM then calculates which nodes share an edge with nodes in S_1 and saves their indices as S_2. In the second iteration, it computes the set of nodes connected to S_2 but eliminates any nodes it has previously visited, and saves the result to S_3. These steps are repeated until the entire system has been explored. For a locally connected graph like a single-orbital tight-binding model, the RCM technique will actually appear as a wave that emanates from the seed until it has filled the entire system.

To give the reader a precise account of RCM, we describe it in terms of the matrix A and its indices in Algorithm 2.

B. Applying RCM on a Model System

We demonstrate how the RCM algorithm would apply to a model system which consists of just six sites arranged in a ring as depicted in Figure 1. This geometry is one of the simplest diversions from a linear topology, and we can write the Hamiltonian for this system as

$$H = \begin{pmatrix} \epsilon_1 & t & 0 & 0 & 0 & t \\ t & \epsilon_2 & t & 0 & 0 & 0 \\ 0 & t & \epsilon_3 & t & 0 & 0 \\ 0 & 0 & t & \epsilon_4 & t & 0 \\ 0 & 0 & 0 & t & \epsilon_5 & t \\ t & 0 & 0 & 0 & t & \epsilon_6 \end{pmatrix}$$

where $\epsilon_{1\ldots6}$ are on-site potentials and t is the hopping element between neighboring sites. If the sites were arranged in a straight line, the Hamiltonian would be trivially block-tridiagonal. But because the sites are now arranged in a ring, the two hopping-terms in the extreme off-diagonals break this property. To compute H^{-1} at site 1, one might naively invert the entire 6×6 matrix, with an associated 6^3 scaling.

To resolve this, RCM begins with a seed index and moves out through the system, keeping track of indices along the way. If we set the seed to site 1, RCM would obtain a series of indices $S_1 = \{1\}, S_2 = \{2, 6\}, S_3 = \{3, 5\}, S_4 = \{4\}$, which allows us to construct a permutation matrix by placing 1's in a zero-matrix. As we move down each row we place a 1 in a column that matches an index in one of the RCM sets, beginning with S_1, then S_2 and so on. The order within each set doesn't matter. For example, the above construction would give us

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

by which we can then compute

$$H' = PHP^T = \begin{pmatrix} \epsilon_1 & t & t & 0 & 0 & 0 \\ t & \epsilon_2 & 0 & t & 0 & 0 \\ t & 0 & \epsilon_6 & 0 & t & 0 \\ 0 & t & 0 & \epsilon_3 & 0 & t \\ 0 & 0 & t & 0 & \epsilon_5 & t \\ 0 & 0 & 0 & t & t & t\epsilon_4 \end{pmatrix}$$

The bandwidth of the Hamiltonian has been reduced, converting it into a 4×4 block-tridiagonal form. The Hamiltonian is then reversed in accordance with Algorithm 2 Step 4 by a final permutation using

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

C. Using RCM to Block-Invert an Open System

Here we explore the role of the seed in RCM. For a given system, the optimization of RCM is entirely determined by this choice. But because block-diagonal inversion using RCM only produces blocks along the diagonal, which entries we need to calculate will also affect our choice. For instance, if only the diagonal entries of the Green's function matrix are needed, then any seed will suffice because these entries will always be returned by the inversion algorithm. We could even sample different seeds to see which would provide the most efficient permutation, although this process could prove computationally expensive. When a transmission calculation is required, however, there is an additional restraint to LRGF: the sub-matrix of the Hamiltonian describing coupling of the system to the environment is a dense matrix because of the contribution from the self-energy. Since we also need to produce all of these entries from the full inversion, it is necessary then to choose this as our seed.

Figure 2: Recursive Green's function methods demonstrated on a nearest-neighbor single-orbital tight-binding Hamiltonian for a graphene stadium. Each block of the Hamiltonian is represented by different colors of atoms. In (a) and (b), we have a common two-terminal left-to-right system. In (a) the recursive algorithm is determined by LRGF, and in (b) by the Outward Wave method. In (c), a single lead enters from the left for studying the full scattering matrix of reflected wavefunctions, as described in Section III E. In (d), the Outward Wave method is applied to an ensemble situation, in which the entire boundary of the device region is treated as a potential location for attaching a lead, as described in Section III D.

The LRGF technique circumvents some of these restrictions by adding an additional step that permits block inversion to produce entries at the extreme diagonal block which conveys information from the input lead to the output lead. RCM could be adapted to this calculation by setting the seed to the input lead, and propagating to the output seed. For any system that deviates from a simple linear topology, however, we run into problems. As soon as a single index in the RCM routine is a member of the output boundary, the fact that the Hamiltonian for the output boundary is dense requires that all remaining indices be contained in the final block, which can result in an unnecessarily large block to invert. In fact, this very limitation was argued by Wimmer and Richter[28] as the motivation for developing their automatic procedure.

We instead propose to set the seed as the collective boundary between the central region and *all* the leads. The first for-loop in Algorithm 1 will then provide us the Green's function at the boundary G_{B}. This block is useful for transmission calculations since it automatically contains the off-diagonal entries corresponding to G_{LR}. In fact, G_B can be permuted as

$$G_B = \begin{pmatrix} G_L & G_{LR} \\ G_{RL} & G_R \end{pmatrix} \tag{10}$$

Since the boundaries between a device and its leads generally lie at the device perimeter, RCM will first search into the interior of the device from the boundary regions at the perimeter. When this search is reversed as in the last line of Algorithm 2, RCM will produce a set of layers that appear to emanate from the interior of the device and radiate toward the leads. In the final step, the waves will converge upon the leads at the same time. We explore several geometries where this happens in Figure 2. Because of the appearance of the set of indices as an outward-moving wave, we call the techniques described in this paper as the Outward Wave Method.

D. Application to Ensembles

For an ensemble of systems where the interior of each system is identical, but only the coupling to the environment changes, we can choose the seed wisely to enhance efficiency over the whole ensemble. This can be accomplished by defining the seed as the set of all possible boundaries in the ensemble. Potential applications include examining the response of a single device to varying lead geometries, which will attach at different points along the device perimeter.

We demonstrate this application by assuming that we have a device where we have set the seed at its entire perimeter. The first for-loop in Algorithm 1 will provide us the Schur complement at the penultimate block Σ^L_{N-1} (see Step 2a of Algorithm 1). For each member in the ensemble, this Schur complement will stay the same while the lead self-energy Σ^r, and thus H'_B, changes. We can then calculate for each member

$$G_B = \left(H'_B - E - \Sigma^L_{N-1}\right)^{-1} \tag{11}$$

which provides us the transmission information by Equation 7. Performing this one inversion over the device perimeter saves considerable computational time over the ensemble, enabling the examination of vast arrays of device-plus-lead ensembles.

E. Extension to the Full Scattering Matrix

In systems in which the entire wavefunction is reflected, there is no distinction between and input and output boundary. While the total reflection coefficient is of course unity at all energies, mixing between the modes can be of scientific interest, for instance, when examining quantum ergodicity (see Kaplan and Heller[12] for an application to the tilted billiard). In this case, there is no alternative available to full matrix inversion, except for sparse matrix routines like SuperLU[8]. For full scattering matrix calculations of this variety, the Outward Wave method contributes an efficient dense-matrix algebra equivalent.

For a scattering matrix that has been block-decomposed into

$$s = \begin{pmatrix} r & t' \\ t & r' \end{pmatrix} \tag{12}$$

it is not sufficient to simply extend the analysis in Section II A, that is, for mode m, $r_{mm} \neq \sqrt{\Gamma_m} G_{mm} \sqrt{\Gamma_m}$. This is because the coupling matrix Γ includes boundary conditions of *both* incoming and outgoing waves. Since there is no obvious numerical solution to calculating an equivalent coupling matrix for only incoming and outgoing waves, we propose the following method.

It is possible to diagonalize the self-energy matrix Σ^r for all boundary conditions projected onto the surface of the device region. The resulting eigenvectors will represent orthogonal modes on the boundary, which can later be converted into asymptotic modes far away from the system. Each one of these orthogonal boundary modes will propagate from the system in separate coherent processes: precisely the basis we are looking for, since this basis will not mix asymptotic modes. Since the boundary region is often quite small compared to the scale of the system, and diagonalization requires on order N^3 operations, the same as inversion, this step contributes little overhead. At this point, it is possible to compute the transmission matrix from any boundary mode to all other boundary modes of the system, since by definition they are all orthogonal. That is, it becomes straightforward to calculate all off-diagonal entries of s. Unitarity of S where $S = ss^\dagger$, can be recovered by imposing that $s_{nn} = 1 - \sum_{m \neq n} |s_{nm}|^2$ so that $\sum_m |s_{nm}|^2 = 1$. An example of this type of calculation is offered in Section IV C.

Could there be a use for a transmission matrix built from $t_{mm} = \sqrt{\Gamma_m} G_{mm} \sqrt{\Gamma_m}$? Yes: this problem maps onto calculating the transmission across an infinite lead, where perpendicular to the lead is attached the equivalent device region. An example of this geometry is depicted in Figure 6. Such geometries are similar to that of a Helmholtz resonator, where current flow is absorbed by the resonator at some energies and enhanced at others, as a result of interference between the direct wavefunction and the wavefunction reflected in the resonator. We use this calculation to contribute to the physical picture of a Fano resonance in Section IV C.

IV. COMPUTATIONAL EXPERIMENTS

We demonstrate the efficacy of our algorithm on a demonstration graphene system: the "relativistic stadium" geometry, which was first explored by Huang et al.[10]. We choose the single-orbital tight-binding model for graphene described in Equation 1 as our basis since it is the current *de facto* standard for computer simulations on graphene of this type (see, for example, Munoz-Rojas et al.[18]) and is the model used in the reference[10].

A. Relativistic Stadium

To validate our code, we compare our transmission results with those of Huang et al. in Figure 3. In addition, we compared our results among full inversion, LRGF, and Outward Wave methods and achieved identical results within machine precision. Compared to the published data, which we have sampled numerically from their article, we find that we achieve nearly-identical results for the system, except near singularities in the density of states, which appear as sharp transmission fluctuations. A close examination reveals that these deviations are numerical artifact partly as a consequence of choosing slightly dif-

Figure 3: Transmission coefficient for the graphene stadium[10] using the Outward Wave method (black) and the original data (grey dashed). Differences between the two data are shown at bottom. Deviations arise from disparities in sampling points and the infinitesimal η parameter. We publish results for a very small η parameter of 2.7×10^{-5}eV.

ferent sampling points in the energy spectrum. Near singularities, even slight differences in where we sample the energy spectrum will have a significant impact on the reported value, making it very difficult to align with the published results exactly. The broader differences, most notably near E=1.938eV and 1.985eV can be accounted for by another numerical artifact: a discrepancy in the size of the infinitesimal η parameter in calculating the self-energies of the leads. Since the value chosen in the original article is not published, and solutions approach an asymptote with smaller η parameters, we have chosen to present our results using a relatively small η parameter of 2.7×10^{-5}eV.

B. Relativistic Stadiums of Various Sizes

For a linear system in which the length of the boundary region is comparable to the width along each segment of the system, LRGF actually offers a factor of 4 improvement in efficiency over the Outward Wave method. Even though there are twice as many sub-matrices to invert in this case, each sub-matrix is now half the size compared to the Outward Wave method, that is, $\sum_{m=1}^{N_L} N_m^3 \to \sum_{m=1}^{2N_L} \left(\frac{1}{2}N_m\right)^3 = \frac{1}{4}\sum_{m=1}^{N_L} N_m^3$. There is a cross-over point, however, where each block in Outward Wave is equal to or smaller than the sub-matrices in an equivalent LRGF calculation. This occurs when the minimum distance between the input and output boundaries, L, satisfies

$$L \leq \frac{N}{2N_B}$$

where N is the number of basis functions in the device and N_B is the number of basis functions along the boundary.

To test this, we created an ensemble of 40 relativistic stadiums. Each has the same radius at the rounded edges of $30a$ where a is the lattice constant of graphene. However, the length along the straight section was varied by a linear function according to the system size parameter. We benchmarked fifty energy points within the spectrum of 1.92 and 2.02 eV using the Harvard Odyssey cluster with dual Xeon E5410 2.3Ghz quad core processors. The results of our benchmarks appear in Figure 4. Most prominently, we find the cross-over point between LRGF and Outward Wave to occur around a system size parameter of 12. For our largest system, we found over a 100-fold improvement for the Outward Wave method over the linear recursive method.

Figure 4: Top: Time of calculation for a single energy point for relativistic stadiums using full inversion (blue stars), LRGF (green crosses), and the Outward Wave method (red pluses). Standard deviations above and below are indicated by whisker bars. Middle: Estimated time for transmission calculations, in arbitrary units, computed from the optimization function discussed in Section III A. Bottom: Estimated memory requirements for each system. The clusters we used had a memory limit of 16GB, which is indicated by the magenta dotted line. Our simulations suggest that the memory estimation for LRGF and the Outward Wave method are undervalued. For the transmission calculation, all three methods returned the same transmission coefficient at the energy point within the precision of the machine. The number of basis functions in each calculation is a linear function of the system size parameter.

We expected the calculation time for full inversion to be the largest of the three methods, and to fail above a certain system size parameter because of memory requirements, which we find in our results above a system size parameter of 25. The reduction in variance is partly explained because memory allocation is a major source of variance in these calculations. All recursive methods require that each sub-block be allocated to memory, and as these blocks grow larger, the relative allocation time also grows (which is shown in the other methods). For nodes with shared memory, interference in this step can be a significant factor. The full inversion method, on the other hand, only requires one allocation. In addition, the load balancer is likely shifting these calculations to nodes with identical processors but different priorities, which suggests the results for full inversion would actually be larger than what we report if all of our simulations ran on identical nodes. Happily, this would open the gap between the methods in terms of efficacy even further.

Above a system size parameter of 19, many of our time trials for the full inversion method failed. As a result, our times show a stark bump in value. To understand this, we estimated the memory requirements for each method by allocating a double-precision complex number for every element of the matrices used. We show our results in Figure 4. The clusters we used had a memory limit of 16 GB, which is indicated by the magenta dotted line. Our predictions are consistent with the bump in time trials for full inversion, since above a system size parameter of 19, the cluster would run out of memory and rely on virtual memory on the hard disk.

In addition, Figure 4 demonstrates the memory benefits of recursive methods in general, but especially the Outward Wave method when a system is large compared to the distance between its input and output boundaries. Memory use becomes especially important considering the memory challenges we faced for full inversion.

Time trials using the LRGF method failed due to memory limitations above a system size parameter of 38, which surprised us since the data themselves wouldn't have breached the memory limit. However, since the

Figure 5: The size of each sub-matrix, in order of sub-matrix index, for Outward Wave (red pluses) and LRGF (green crosses) for stadiums of system size parameters 40 and 10 (insert).

recursive algorithms require allocating many blocks of memory of varying sizes, it is very likely that the pointer tables and the allocation process induce memory overheads.

We also modeled the estimated time of calculation for standard linear recursive and outward wave methods using the optimization function $O(\{N_i\}) = \sum_i N_i^3$ and found the same cross-over point at system size parameter 19 (Figure 4). For our ensembles, we found it difficult to determine whether the time of inversion or the challenges with allocating and storing memory were the dominant factors in the final calculation times. We did not plot the equivalent results using full inversion since the underlying algorithm is different.

To understand how each method contributed to the optimization function, we also plot the size of each matrix that must be inverted for the LRGF and Outward Wave methods for stadia of system size parameter 10 and 40 in Figure 5 as in Wimmer and Richter[28]. The area underneath each function is the same and adds to the total number of orbitals in each system. As a result, each curve represents, in effect, the bandwidth of the sparse Hamiltonian according to the two permutations. The better the permutation, the smaller the overall bandwidth the shorter (and wider) it will appear in this graphic. At the system size parameter 10, which is near the cross-over point at 12, we find very similar matrix bandwidths for the two methods, which corroborates both our predicted and measured calculation times. Beyond the cross-over point, the Outward Wave method requires the inversion of many more matrices but of far smaller size, giving an overall performance boost.

C. Reflection Matrix For Single-Lead Relativistic Stadium

We choose to examine the single transmission fluctuation at $E = 1.9584$eV in Figure 3. The physical explanation for such transmission fluctuations is well accounted for by Fano resonance theory[9] which provides a succinct formula that models the conductance fluctuation as

$$G(\epsilon) \propto \frac{(\epsilon + q)^2}{\epsilon^2 + 1} \tag{13}$$

Here ϵ is the energy of the system, zeroed at the center of the resonance, and q is an asymmetry factor. Fano proposed that these conductance fluctuations result from the interference of a directly and an indirectly (resonant) scattering state. This theory suggests that the breadth of the resonance (and the conductance fluctuation) will be proportional to the coupling between the resonant mode and the environment (leads), and that the asymmetrical q-factor can be accounted by the relative phase between the directly scattering state and the indirectly scattering state. We can test these implications by comparing the two-lead stadium to an equivalent simulation where the incoming and outgoing leads are in fact an infinite nanoribbon with a stadium resonator attached perpendicular to the direction of flow, as depicted in Figure 6. This scenario can be described as a Helmholtz resonator as discussed in Section III E.

Figure 6: Geometry for the Helmholtz resonator configuration.

We expect three changes to happen for the Helmholtz resonator:

1. The energy of the resonance, and thus the center of the conductance fluctuation, will shift to reflect the change in coupling matrix.

2. The resonance width will reduce by a factor of two, to reflect that we have reduced coupling between the resonant state and the environment by half.

3. It has been suggested by Racec et al.[22] that the asymmetry q-factor can be explained by the relative lateral symmetry between the direct and scattering states. In this case, we expect the asymmetric pattern in transmission to reverse, to reflect the fact that we are now reflecting off the same side of the system, as opposed to tunneling through it. From this perspective, our calculation is an excellent validation of Racec's study.

Each point is beautifully verified in Figure 7. For instance, the peak at $E = 1.9584$eV shifts down by 0.0003eV and its resonance width is divided by a large fraction. In fact, it is much smaller than we predicted and suggests that there may be additional factors constricting the resonant width in the Helmholtz resonator scenario. In addition, we see that its asymmetric profile has reversed, reaching a transmission minimum before its transmission maximum. In both cases, the transmission fluctuation traverses approximately one unit.

In addition, we computed the full reflection matrix of the single-lead stadium equivalent as depicted in Figure 2c, using the method described in Section III E to elucidate the role of mode-mixing to the Helmholtz resonator transmission function. There are exactly three open modes in this energy range, so that a similar calculation with just the nanoribbon would give a flat profile at a transmission coefficient of 3. With the addition of the Helmholtz resonator, one of those modes is fully reflected and the remaining two are partially reflected and mix. We perform eigenchannel analysis on the remaining two modes and find a small amount of mixing which varies in proportion to the background slope of the transmission function, indicating mode-mixing as a

Figure 7: Transmission function for the relativistic stadium (dashed), its Helmholtz-resonator equivalent (solid), and mode-mixing in the single-lead reflection matrix (dotted).

System Size Parameter	SuperLU	Outward Wave	Ratio
80	57.3	49.5	1.158
100	76.7	56.3	1.362
200	156.3	95.1	1.644
300	247.6	132.0	1.876
400	340.8	174.8	1.950

Table I: Comparison of time trials (in average seconds per energy trial) for rectangular graphene stadia. As the system size parameter increased, so did the improvement in efficiency for Outward Wave over SuperLU.

salient feature of either eigenchannel. If either of these channels is coupled to a resonant state whose probability function peaks at a resonant energy, the channel will strongly couple to the environment at that energy, and we should find this reflected as a small but noticeable peak in the amount of mode-mixing at the same energy. Using the one-lead reflection matrix, we find precisely such a peak of mode-mixing at the resonant energy of the Helmholtz resonator, as reflected in Figure 7.

D. Sparse Matrix Packages

We performed a set of experiments using a variation of the relativistic stadium with square ends. We tested our time trials using our Outward Wave method against an equivalent calculations using sparse matrix inversion for the required elements using SuperLU[8]. In each trial, we kept the length of the system identical, but increased its width according to the system size parameter, as with the relativistic stadium trials. We report the results in Table I. In all experiments, we obtained identical results for transmission to within precision of the machine. For both algorithms, we found a similar scaling of computation time with the system size parameter. We found that for small systems, both algorithms returned results in approximately the same time scale. As the systems grew larger compared to the distance between the input and output boundaries, however, we found efficiency gain for Outward Wave, approaching a factor of two for our largest system. We attribute the efficiency gain of our code to the fact that the algorithm and software are specifically tailored to our problem. Moreover, the roughly equivalent scaling with system size between the Outward Wave method and SuperLU corroborates that Outward Wave achieves a close-to-optimal block-diagonalization of the Hamiltonian, and comes closest to the ideal case for systems that are large compared to the shortest path between the input and output boundaries. This would be the case, for instance, when the input and output boundaries overlap, as in Section III E.

V. CONCLUSIONS

We have shown an alternate perspective of the linear recursive Green's function method for transmission and LDOS calculations that moves from a left-to-right paradigm to an interior-to-exterior paradigm. The new perspective, which we dub the Outward Wave method, permutes the Hamiltonian into a sequence of blocks which begin at the interior of the device and progress toward the leads. The Outward Wave method works entirely from the Hamiltonian and the leads' self-energy and allows one to enjoy the computational scaling of the linear method while expanding the geometries available to their calculations.

In addition, we have shown that this perspective, along with considerations of a proper basis set for the boundary, can be used to efficiently calculate the entire full-spectrum scattering matrix for any system. We demonstrate the power of such a tool to contribute to a physical picture behind the Fano resonance in a relativistic stadium, corroborating the studies of Huang[10] and Racec[22]. It is our hope to use this tool to examine the mixing of reflected modes in weakly ergodic systems in future studies.

Finally, we have compared our results to a state-of-the-art sparse matrix package and found similar scaling by system size, corroborating the efficiency and general applicability of our algorithm.

Acknowledgments

This research was conducted with funding from the Department of Energy Computer Science Graduate Fellowship program under Contract No. DE-FG02-97ER25308. Code and development for this work, as well as computer time for the SuperLU simulations, were conducted with D. P. and J. N. as a User project at the Molecular Foundry, Lawrence Berkeley National Laboratory, which is supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

[1] H. U. Baranger, D. P. DiVincenzo, R. A. Jalabert, and A. D. Stone. Classical and quantum ballistic-transport anomalies in microjunctions. *Phys. Rev. B*, 44(19):10637–10675, Nov 1991.

[2] S. Cauley, J. Jain, C.-K. Koh, and V. Balakrishnan. A scalable distributed method for quantum-scale device simulation. *Journal of Applied Physics*, 101(123715), 2007.

[3] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, ACM '69, pages 157–172, New York, NY, USA, 1969. ACM.

[4] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172, New York, NY, USA, 1969. ACM.

[5] S. Datta. *Electronic Transport in Mesoscopic Systems.* Cambridge University Press, Cambridge, 1997.

[6] T. A. Davis. *Direct Methods for Sparse Linear Systems.* SIAM, 2006.

[7] J. W. Demmel. *Applied Numerical Linear Algebra.* SIAM, 1997.

[8] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Analysis and Applications*, 20(3):720–755, 1999.

[9] U. Fano. Effects of configuration interaction on intensities and phase shifts. *Phys. Rev.*, 124(6):1866–1878, Dec 1961.

[10] L. Huang, Y.-C. Lai, D. Ferry, S. Goodnick, and R. Akis. Relativistic Quantum Scars. *Physical Review Letters*, 103(5):1–4, July 2009.

[11] B. Huard, J. A. Sulpizio, N. Stander, K. Todd, B. Yang, and D. Goldhaber-Gordon. Transport measurements across a tunable potential barrier in graphene. *Physical Review Letters*, 98(236803), 2007.

[12] L. Kaplan and E. J. Heller. Weak quantum ergodicity. *Physica D: Nonlinear Phenomena*, 121(1-2):1 – 18, 1998.

[13] M. P. López-Sancho and J. Rubio. Quick iterative scheme for the calculation of transfer matrices: application to mo (100). *J. Phys. F.: Met. Phys.*, 14:1205–1215, 1984.

[14] M. P. López-Sancho and J. Rubio. Highly convergent schemes for the calculation of bulk and surface green functions. *J. Phys. F.: Met. Phys.*, 15:851–858, 1985.

[15] D. Mamaluy, M. Sabathil, and P. Vogl. Efficient method for the calculation of ballistic quantum transport. *Journal of Applied Physics*, 93(8):4628–4633, 2003.

[16] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl. Contact block reduction method for ballistic transport and carrier densities of open nanostructures, Phys. *Rev. B*, 71(24):245321, 2005.

[17] Y. Meir and N. S. Wingreen. Landauer formula for the current through an interacting electron region. *Phys. Rev. Lett.*, 68(16):2512–2515, Apr 1992.

[18] F. Muñoz Rojas, D. Jacob, J. Fernández-Rossier, and J. Palacios. Coherent transport in graphene nanoconstrictions. *Physical Review B*, 74(19):1–8, 2006.

[19] F. OuYang, J. Xiao, R. Guo, H. Zhang, and H. Xu. Transport properties of t-shaped and crossed junctions based on graphene nanoribbons. *Nanotechnology*, 20(5):055202, 2009.

[20] B. Ozyilmax, P. Jarillo-Herrero, D. Efetov, and P. Kim. Electronic transport in locally gated graphene nanoconstrictions. *Applied Physics Letters*, 91(192107), 2007.

[21] Z. Qiao and J. Wang. A variant transfer matrix method suitable for transport through multi-probe systems. *Nanotechnology*, 18(43):435402, 2007.

[22] E. R. Racec, U. Wulf, and P. N. Racec. Fano regime of transport through open quantum dots. *Phys. Rev. B*, 82(8):085313, Aug 2010.

[23] F. Sols, F. Guinea, and A. H. C. Neto. Coulomb blockade in graphene nanoribbons. *Physical Review Letters*, 99(166803), 2007.

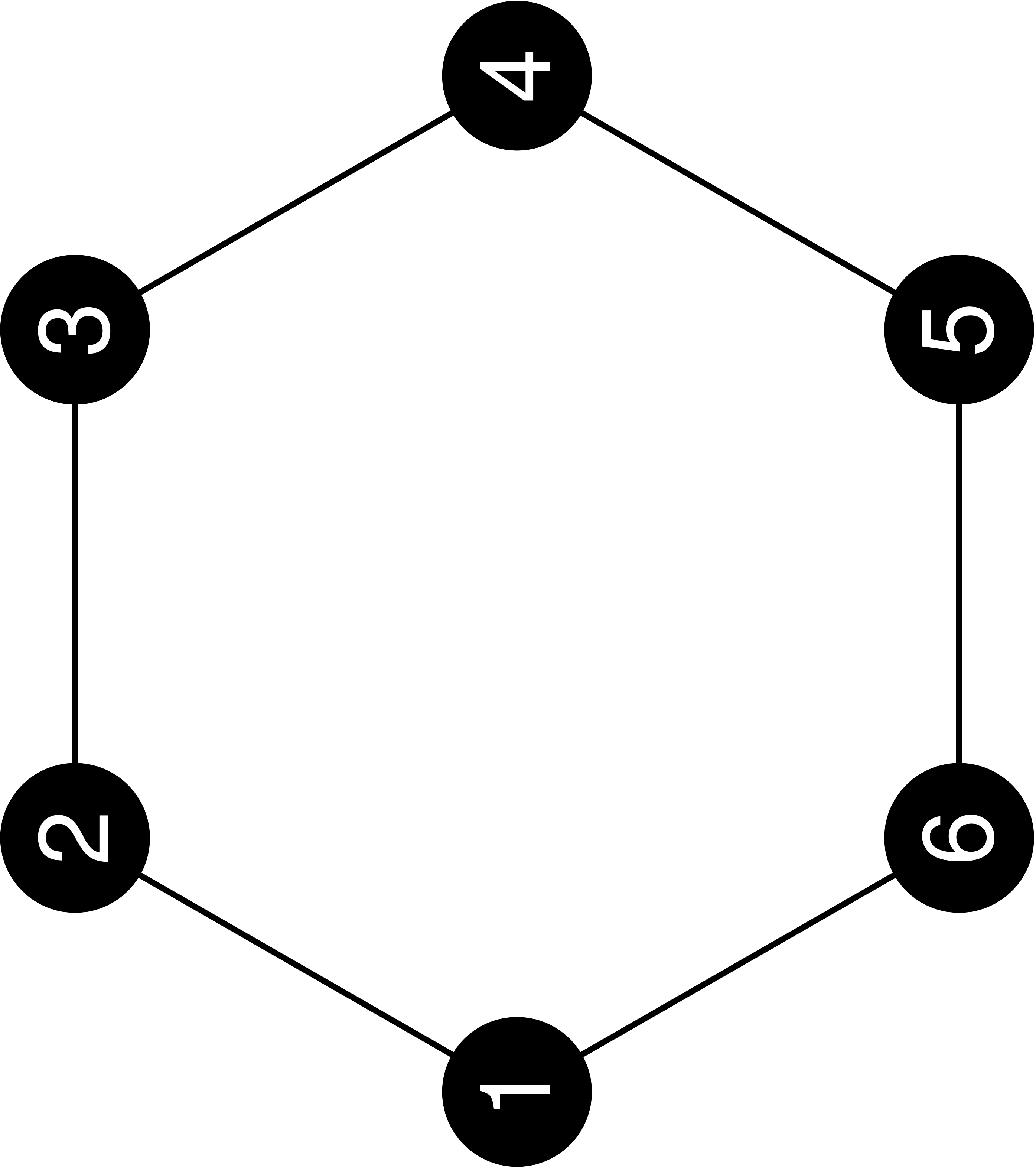
[24] A. Svizhenko, M. P. Anamtram, T. R. Govindand, B. Biegel, and R. Venugopal. Two-dimensional quantum mechanical modeling of nanotransistors. *Journal of Applied Physics*, 91(4):2343, 2002.

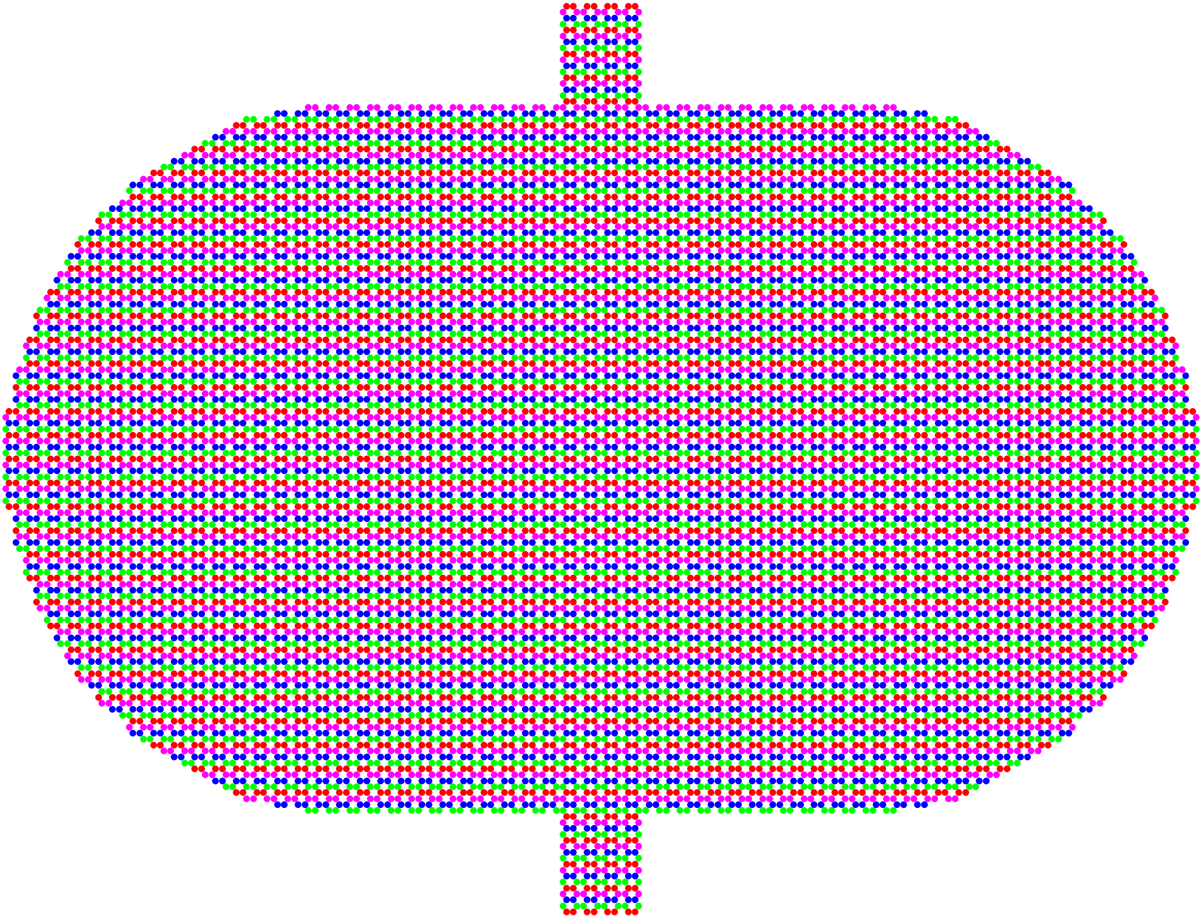
[25] Y.-W. Tan, Y. Zhang, K. Bolotin, Y. Zhao, S. Adam, E. H. Hwang, S. D. Sarma, H. L. Stormer, and P. Kim. Measurement of scattering rate and minimum conductivity in graphene. *Physical Review Letters*, 99(246803), 2007.

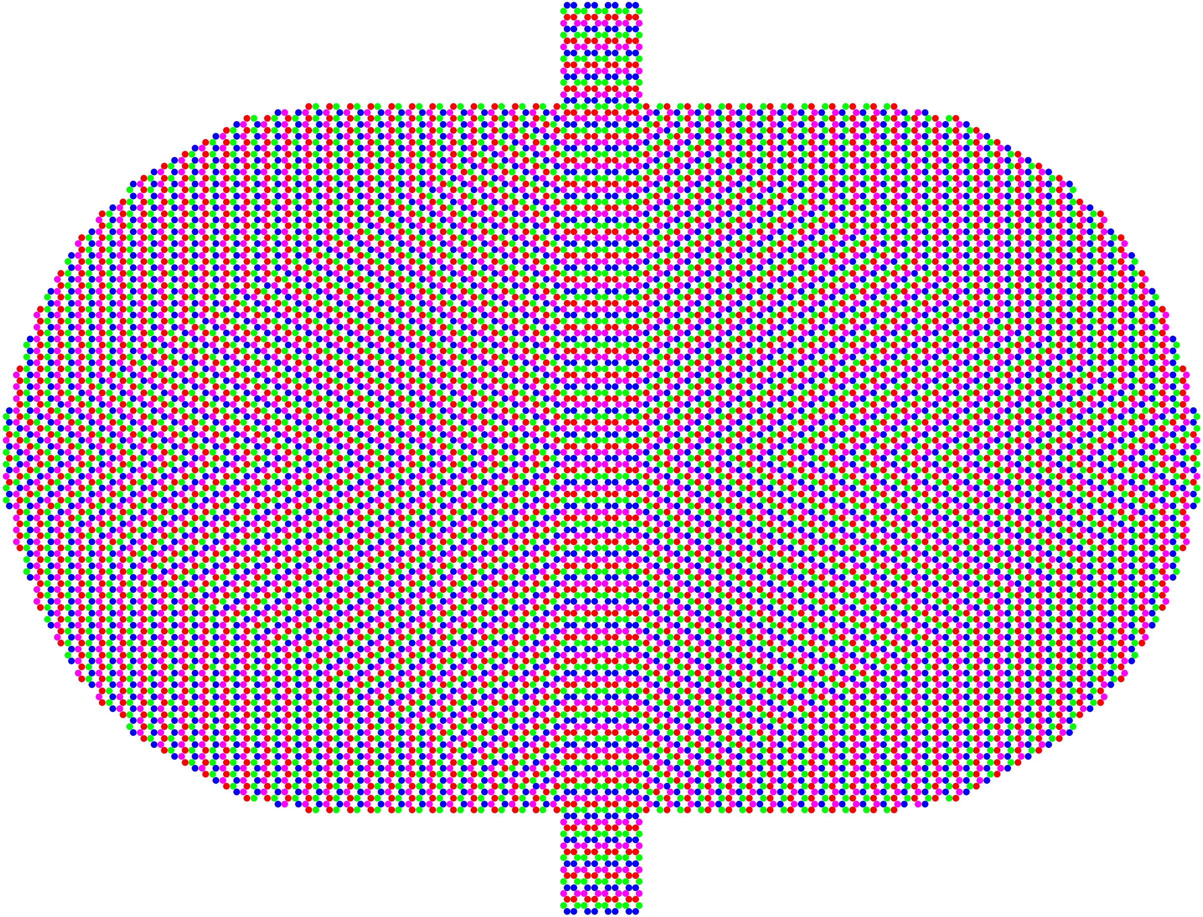
[26] T. Usuki, M. Takatsu, R. A. Kiehl, and N. Yokoyama. Numerical analysis of electron-wave detection by a wedge-shaped point contact. 1994.

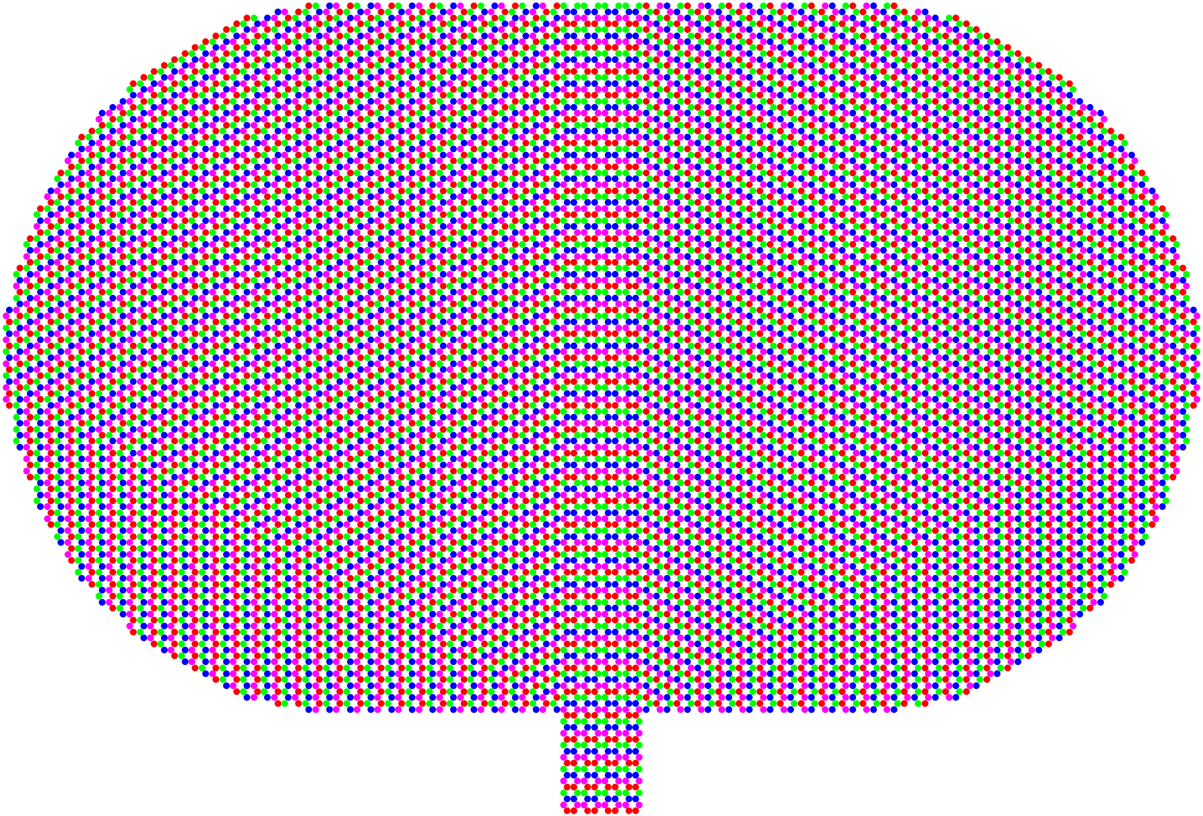
[27] W. L. Wang, O. V. Yazyev, S. Meng, and E. Kaxiras. Topological frustration in graphene nanoflakes: Magnetic order and spin logic devices. *Phys. Rev. Lett.*, 102(15):157201, Apr 2009.

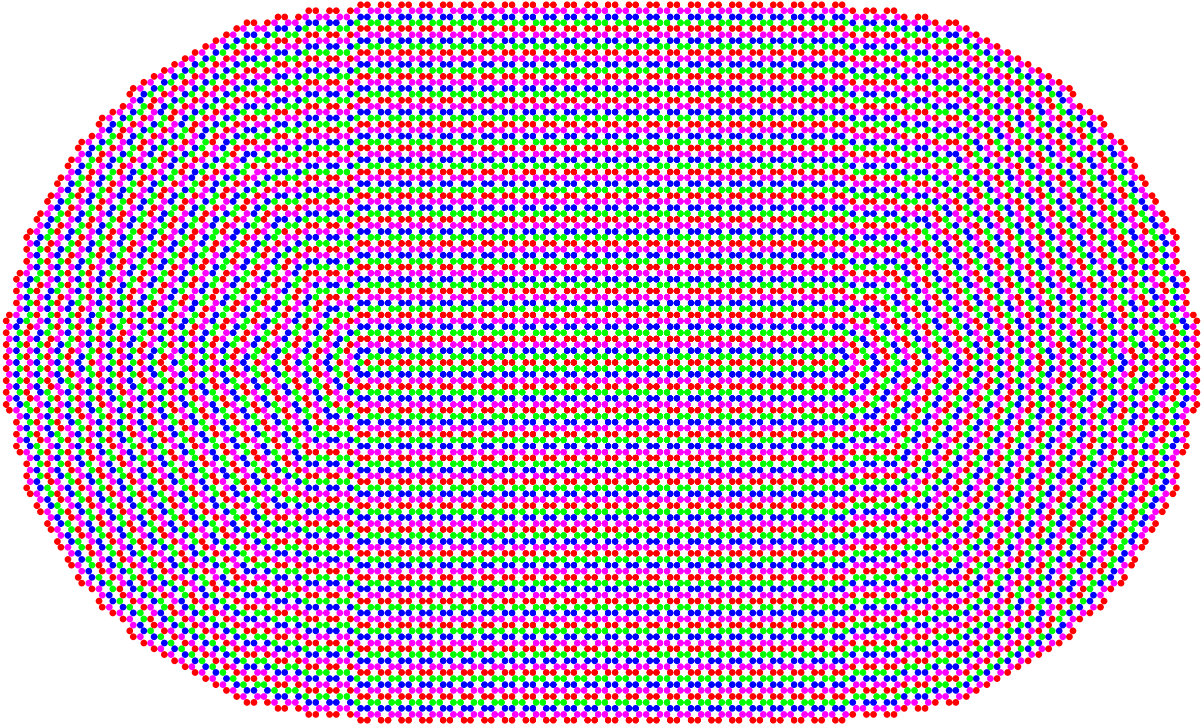
[28] M. Wimmer and K. Richter. Optimal block-tridiagonalization of matrices for coherent charge transport. *Journal of Computational Physics*, 228(23):8548–8565, 2009.

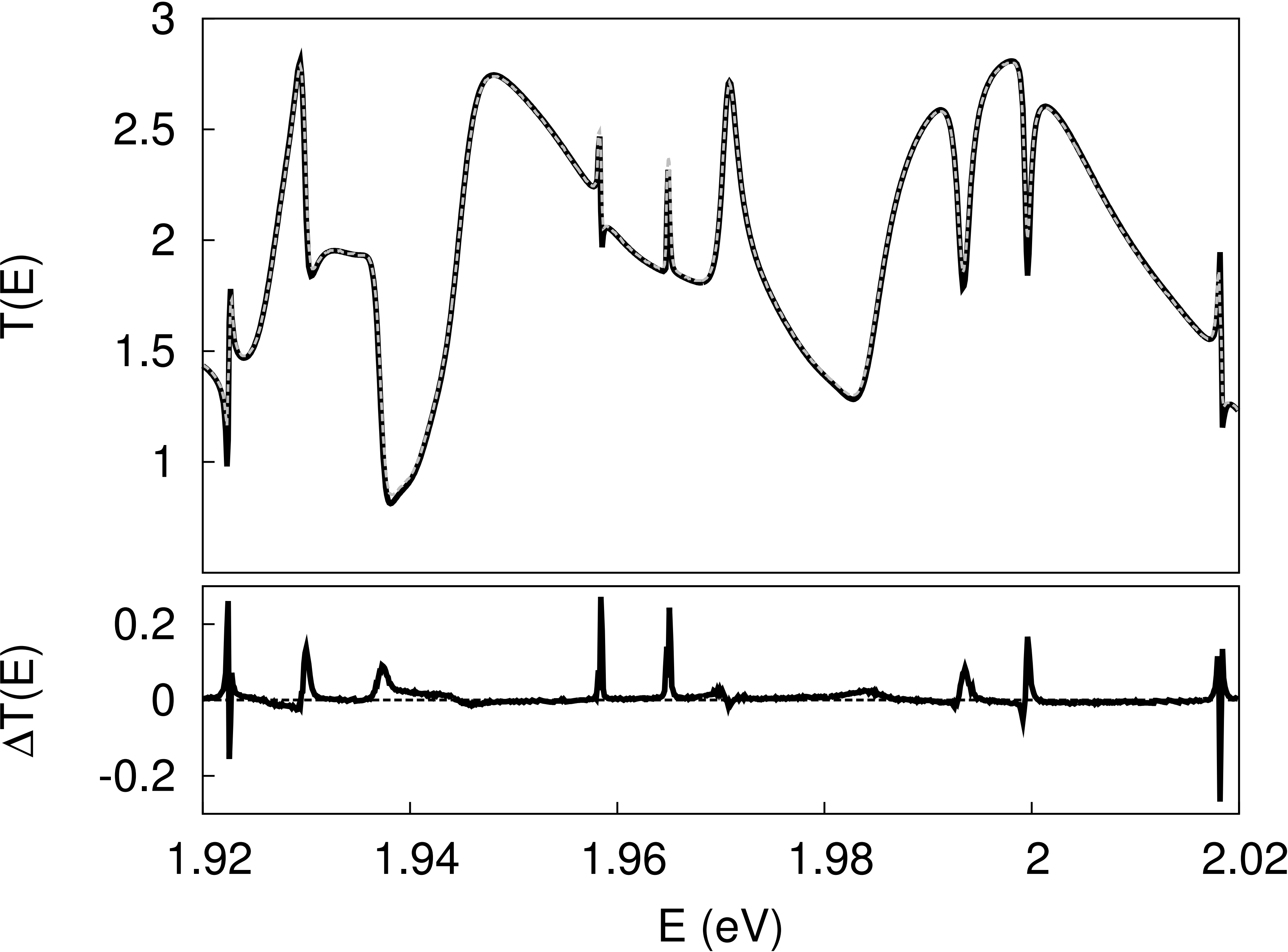

4
3
5
2
6
1

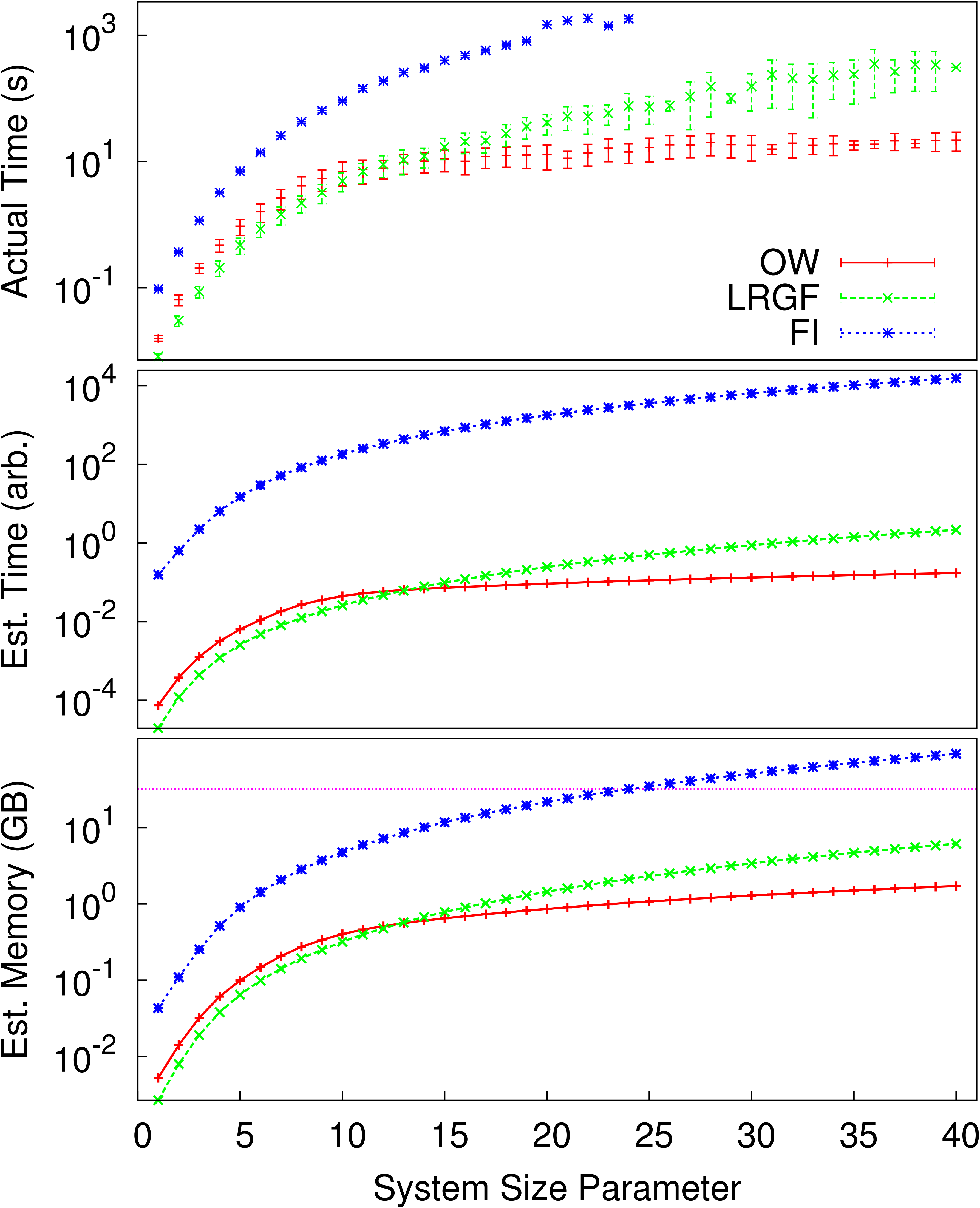


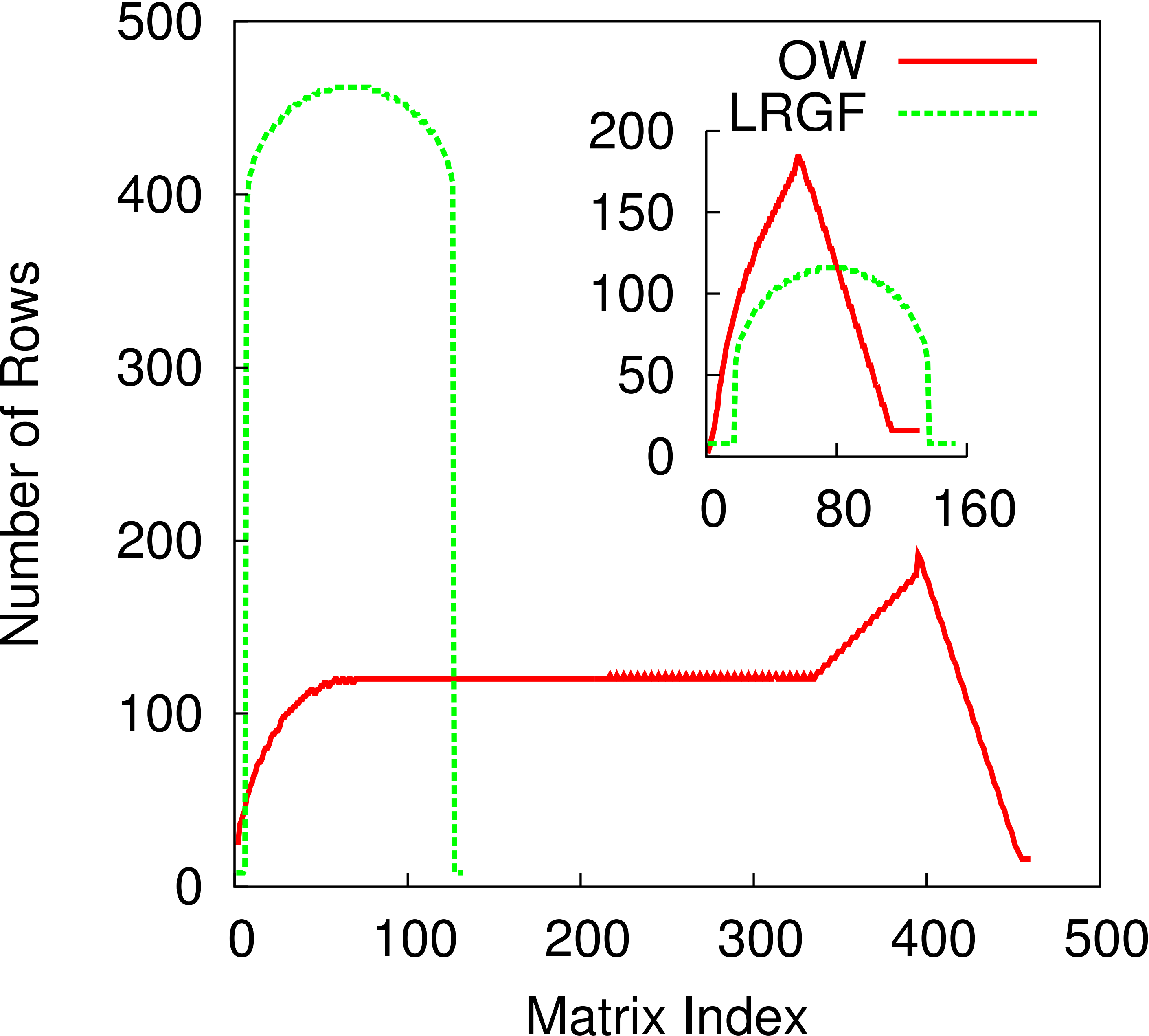


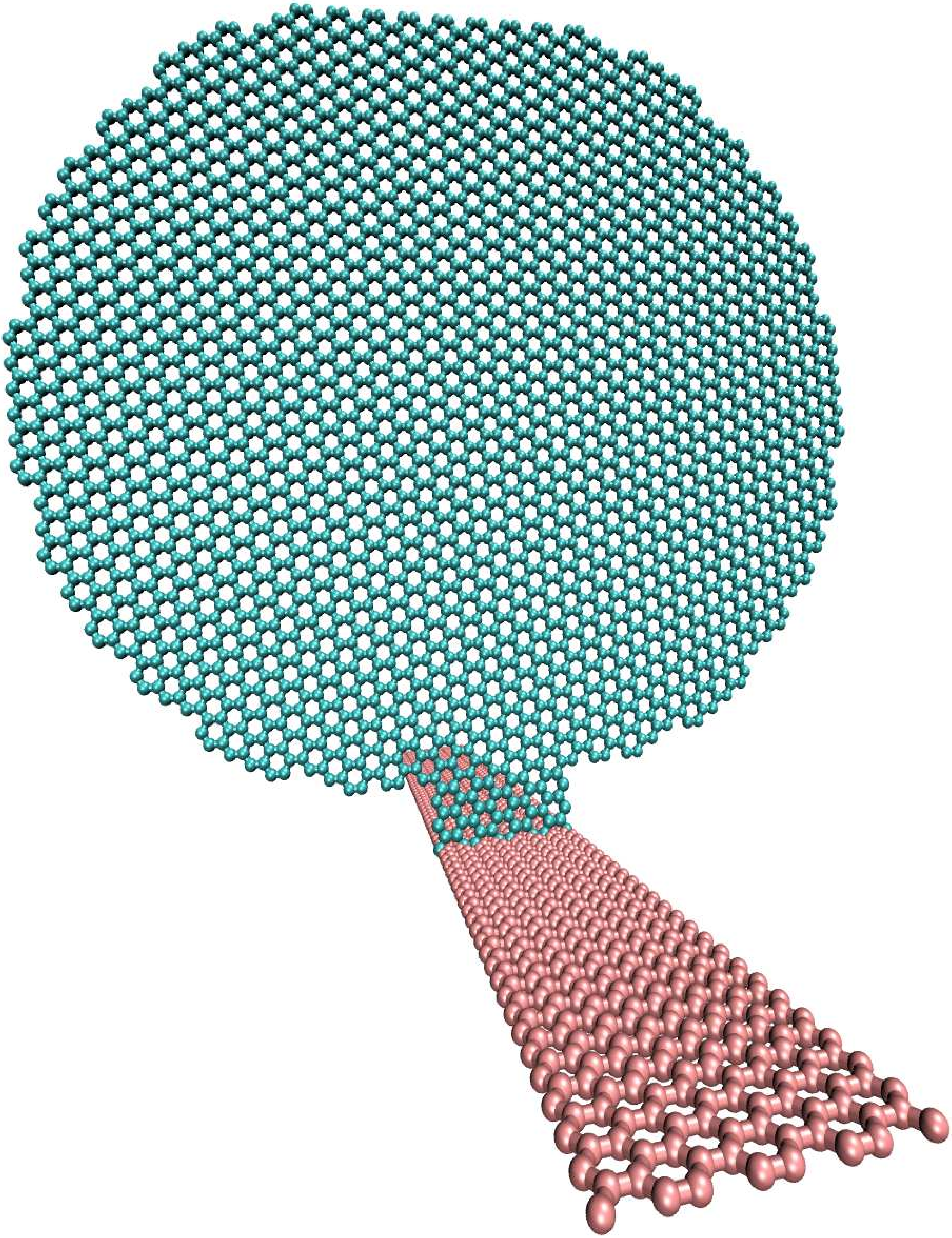


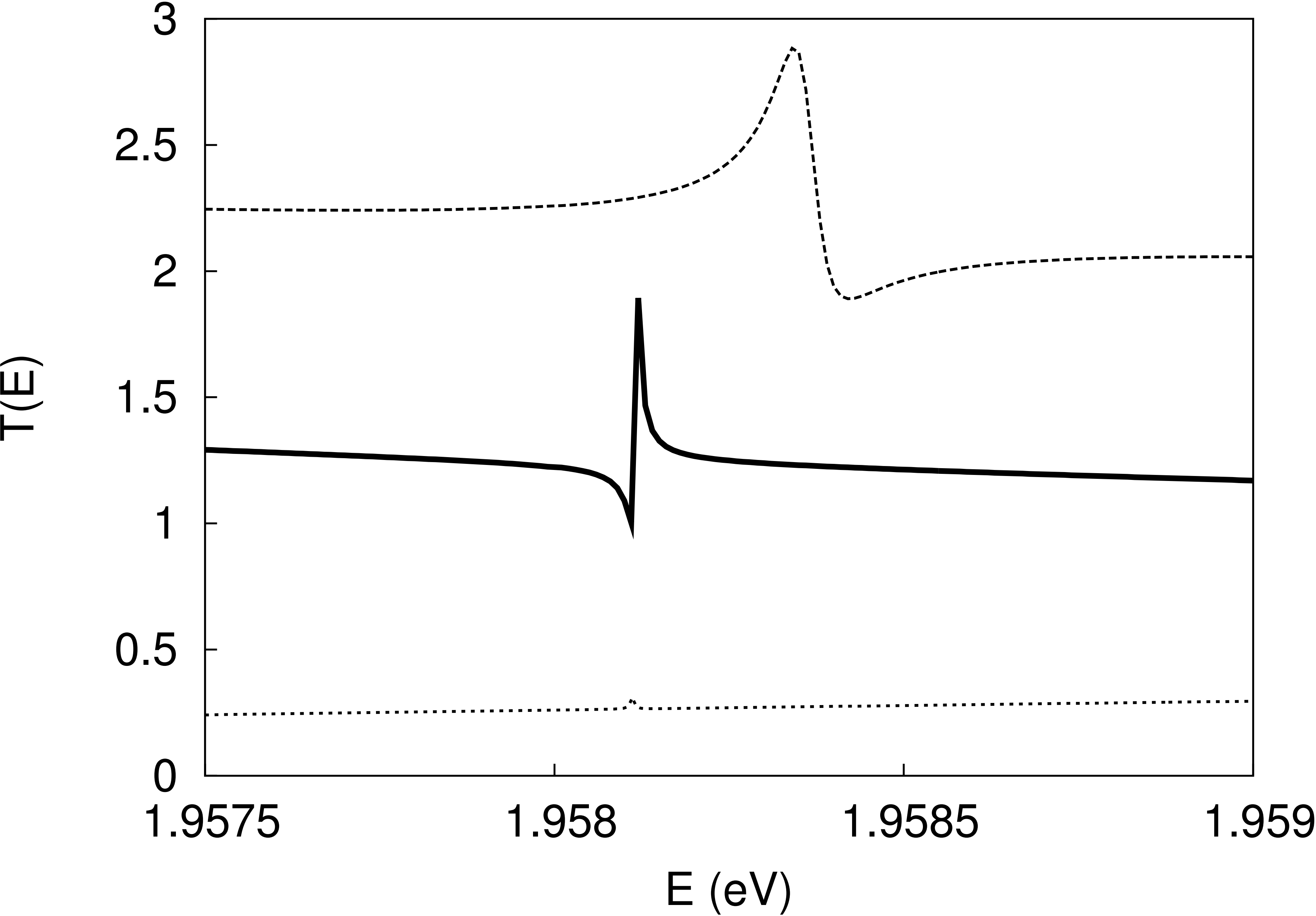



T(E)
3
2.5
2
1.5
1
ΔT(E)
0.2
0
-0.2
1.92
1.94
1.96
1.98
2
2.02
E (eV)


Actual Time (s)
10^3
10^1
10^-1
OW
LRGF
FI
Est. Time (arb.)
10^4
10^2
10^0
10^-2
10^-4
Est. Memory (GB)
10^1
10^0
10^-1
10^-2
0
5
10
15
20
25
30
35
40
System Size Parameter


OW
LRGF
Number of Rows
0
100
200
300
400
500
Matrix Index
0
100
200
300
400
500
200
150
100
50
0
0
80
160




3
2.5
2
1.5
1
0.5
0
T(E)
1.9575
1.958
1.9585
1.959
E (eV)