Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 227 (2008) 3174–3190
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Block tridiagonal matrix inversion and fast transmission calculations

Dan Erik Petersen [a,*], Hans Henrik B. Sørensen [b], Per Christian Hansen [b], Stig Skelboe [a], Kurt Stokbro [c]

[a] *Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark*
[b] *Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Bldg. 321, DK-2800 Lyngby, Denmark*
[c] *Nanoscience Center, University of Copenhagen, Universitetsparken 5d, DK-2100 Copenhagen, Denmark*

Received 24 April 2007; received in revised form 15 November 2007; accepted 19 November 2007
Available online 8 December 2007

Abstract

A method for the inversion of block tridiagonal matrices encountered in electronic structure calculations is developed, with the goal of efficiently determining the matrices involved in the Fisher–Lee relation for the calculation of electron transmission coefficients. The new method leads to faster transmission calculations compared to traditional methods, as well as freedom in choosing alternate Green's function matrix blocks for transmission calculations. The new method also lends itself to calculation of the tridiagonal part of the Green's function matrix. The effect of inaccuracies in the electrode self-energies on the transmission coefficient is analyzed and reveals that the new algorithm is potentially more stable towards such inaccuracies.

PACS: 71.15.−m; 02.70.−c

Keywords: Matrix inversion; Electron transport; Transmission; Density functional theory

1. Introduction

Quantum transport simulations have become an important theoretical tool for investigating the electrical properties of nanoscale systems, both in the semi-empirical approach [1–4] and full ab initio approach [5–8]. The basis for the approach is the Landauer–Büttiker model of coherent transport, where the electrical properties of a nanoscale constriction is described by the transmission coefficients of a number of one-electron states propagating coherently through the constriction. The approach has been used successfully to describe the electrical properties of a wide range of nanoscale systems, including atomic wires, molecules and interfaces

* Corresponding author. Tel.: +45 35 32 14 00; fax: +45 35 32 14 01.
E-mail address: danerik@diku.dk (D.E. Petersen).

0021-9991/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.11.035

[9–18]. In order to apply the method to semiconductor device simulation, it is necessary to handle systems comprising millions of atoms, and this will require new, efficient algorithms for calculating the transmission coefficient.

In this paper, ideas and calculations behind an algorithm that provides an improvement over a widely popular technique employed in the calculation of transmission coefficient of so-called two-probe systems [15] is presented. A two-probe system consists of three regions: a left electrode region, a central scattering region and a right electrode region. The electrode regions are semi-infinite periodic systems, and the scattering region connects the two electrode regions. A one-electron tight-binding Hamiltonian is used to describe the electronic structure of the system. The tight-binding Hamiltonian can be obtained from a semi-empirical tight-binding description as obtained from an extended Hückel model [19] or through a first-principles approach as obtained when using a self-consistent density-functional Kohn–Sham Hamiltonian [20].

In the pursuit of determining the electronic structure of molecules, bulk crystals and two-probe systems, associated self-consistent DFT calculations, relevant Green's functions and ultimately calculation of the transmission of two-probe systems all involve the problem of matrix inversion in some form or another. This paper deals with matrices of a *block tridiagonal* form, which lie at the center of the problems to be solved. Block matrices will be denoted with uppercase bold letters, while lower case bold letters refer to sub-block matrices of their uppercase counterparts.

Throughout this paper, it is assumed that block tridiagonal matrix, $\mathbf{A}$, is dealt with and that it is to be inverted in order to obtain the Green's function matrix (or a part thereof). In the process of finding the Green's function matrix $\mathbf{G} = \mathbf{A}^{-1}$ that enters in DFT theory, the following equation sets up the problem [21]:

$$\mathbf{A} = \varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{L}} - \mathbf{\Sigma}^{\mathrm{R}}. \tag{1}$$

In the above expression $\mathbf{S}$ is an overlap matrix, $\mathbf{H}$ is the Hamiltonian of the system and $\mathbf{\Sigma}^{\mathrm{L}}$ and $\mathbf{\Sigma}^{\mathrm{R}}$ are the self-energies from the *left* and *right* semi-infinite electrodes, respectively. Furthermore, the matrix $\mathbf{G}$ depends on the variable ε that dictates the energy of an incoming one-electron coherent wave for which it is desired to investigate the transmission through the system. The methods developed in this paper are designed for a fixed value of ε.

The individual blocks of the matrix $\mathbf{A}$ are denoted $\mathbf{a}_{ij}$ and are assumed to be dense, complex matrices along the tridiagonal. The diagonal blocks are square matrices, while the off-diagonal blocks are typically rectangular. The structure of $\mathbf{A}$ for two relevant cases is shown in Figs. 1 and 2.

A method to obtain the Green's function matrix $\mathbf{G}$ is now devised, much in the same spirit as [22]. In order to do so, the matrix to be inverted, $\mathbf{A}$, is augmented with the identity matrix, $\mathbf{I}$.

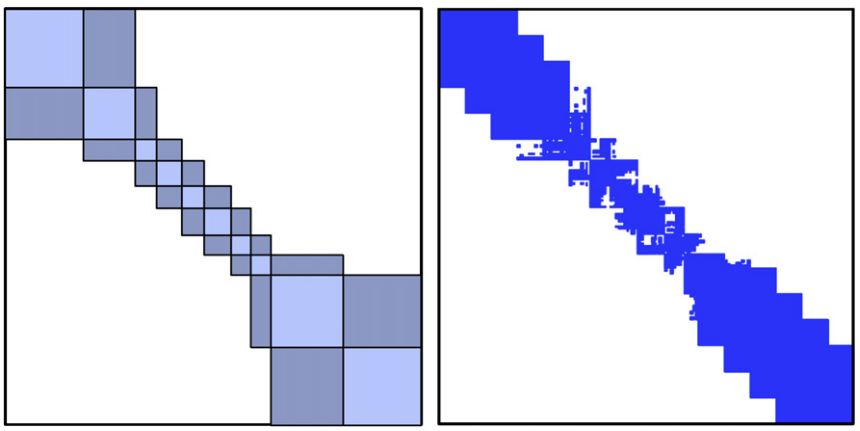

Fig. 1. The block-tridiagonal and sparsity structure for the Au111–AR example [17]. The matrix is of dimension 1295×1295, split up along the diagonal in blocks of order 243, 162, 66, 79, 69, 84, 62, 62, 225, and 243 from upper left to lower right, along with corresponding off-diagonal blocks.

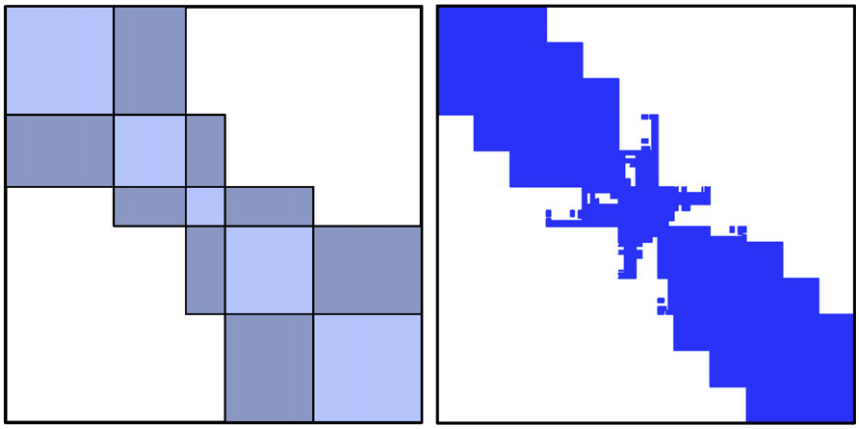

Fig. 2. The block-tridiagonal and sparsity structure for the Au111–DTB example [18]. The matrix is of dimension 943×943, split up along the diagonal in blocks of order 243, 162, 88, 198 and 243 from upper left to lower right, along with corresponding off-diagonal blocks.

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{I} \end{array}\right] = \left(\begin{array}{ccccc|ccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{a}_{23} & & & & \mathbf{i}_{22} & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} \\ & & \ddots & \ddots & \ddots & & & \ddots \end{array}\right) \tag{2}$$

Each diagonal block of the identity matrix, $\mathbf{i}_{ii}$ has the same square block size of the corresponding block $\mathbf{a}_{ii}$ of the matrix $\mathbf{A}$, and are themselves identity matrices.

The organization and shape of the matrix blocks in $\mathbf{A}$ are related to the topology of the two probe system. Looking at, e.g. Fig. 1, portions of the electrodes can be identified as the regions comprised of larger blocks towards the corner of the matrix, while the more sparsely populated central region of the system is identified as a series of smaller matrix blocks in the center of $\mathbf{A}$. The top left corner of $\mathbf{A}$ attaches to the left electrode, while the lower right corner attaches to the right electrode.

The expression *augmented matrix* $[\mathbf{A}|\mathbf{I}]$ is equivalent to the equation $\mathbf{AG} = \mathbf{I}$ (cf. [23]). By manipulating the augmented matrix through a series of operations such that the left side, $\mathbf{A}$, is reduced to the identity matrix $\mathbf{I}$, we will obtain the augmented matrix $[\mathbf{I}|\mathbf{G}]$ where the inverse of $\mathbf{A}$, namely $\mathbf{G} = \mathbf{A}^{-1}$, can be read on the right. This is done by illustrating the forward and backward block Gaussian elimination steps, and then combining the results.

Calculating all of $\mathbf{G}$ is ultimately not of interest. Only a block $\mathbf{g}_{ij}$ of $\mathbf{G}$ to be used in further transmission calculations will be determined. It is the particular choice of $\mathbf{g}_{ij}$ and the procedure for its calculation that separates the new transmission calculation method from previous algorithms.

This paper is organized as follows. The notation and block Gaussian elimination technique on which the methods used in this paper is based on is described in Section 2. Section 3 shows how the result of block Gaussian elimination is used to generate the Green's function matrix $\mathbf{G}$. In Section 4, the calculation of transmission values via a traditional method and a new method is explained. The new method is then benchmarked against the traditional, baseline method, via a consideration of computational complexity, as well as measured speedup times in Section 5. The effects of perturbed surface Green's function matrices on the transmission accuracy, and conclusions on which portions of $\mathbf{G}$ would lead to more accurate transmission calculations is considered in Section 6. Conclusions are finally presented in Section 7.

2. Forward and backward block Gaussian elimination

The forward procedure is characterized with the superscript L since the elimination procedure proceeds from the *left* electrode towards the right.

A block Gaussian elimination step is performed on the matrix given in Eq. (2) by multiplying the first block row by the matrix $\mathbf{c}_1^{\mathrm{L}} = -\mathbf{a}_{21}\mathbf{a}_{11}^{-1}$ and subsequently adding it to the second block row. This produces a zero block in the (2,1) position:

$$\left(\begin{array}{ccccc|ccccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & & & \\ \mathbf{a}_{21}+\mathbf{c}_1^{\mathrm{L}}\mathbf{a}_{11} & \mathbf{a}_{22}+\mathbf{c}_1^{\mathrm{L}}\mathbf{a}_{12} & \mathbf{a}_{23} & & & \mathbf{c}_1^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{i}_{22} & & & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} & & \\ & & \ddots & \ddots & \ddots & & & & \ddots & \end{array}\right) = \left(\begin{array}{ccccc|ccccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & & & \\ \mathbf{0} & \mathbf{a}_{22}-\mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{a}_{12} & \mathbf{a}_{23} & & & \mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{i}_{11} & \mathbf{i}_{22} & & & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} & & \\ & & \ddots & \ddots & \ddots & & & & \ddots & \end{array}\right). \tag{3}$$

Next, a block Gaussian elimination step is performed by multiplying the second row by the factor $\mathbf{c}_2^{\mathrm{L}} = -\mathbf{a}_{32}(\mathbf{a}_{22}-\mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{a}_{12})^{-1}$ and subsequently adding it to the third row. This produces a zero block in the (3,2) position. A recursive routine that will complete a full forward block Gaussian elimination is now defined.

$$\begin{aligned} \mathbf{d}_{11}^{\mathrm{L}} &= \mathbf{a}_{11} & \mathbf{c}_1^{\mathrm{L}} &= -\mathbf{a}_{21}(\mathbf{d}_{11}^{\mathrm{L}})^{-1} \\ \mathbf{d}_{22}^{\mathrm{L}} &= \mathbf{a}_{22}-\mathbf{a}_{21}(\mathbf{d}_{11}^{\mathrm{L}})^{-1}\mathbf{a}_{12} & \mathbf{c}_2^{\mathrm{L}} &= -\mathbf{a}_{32}(\mathbf{d}_{22}^{\mathrm{L}})^{-1} \\ \mathbf{d}_{33}^{\mathrm{L}} &= \mathbf{a}_{33}-\mathbf{a}_{32}(\mathbf{d}_{22}^{\mathrm{L}})^{-1}\mathbf{a}_{23} & \mathbf{c}_3^{\mathrm{L}} &= -\mathbf{a}_{43}(\mathbf{d}_{33}^{\mathrm{L}})^{-1} \\ &\vdots & &\vdots \\ \mathbf{d}_{ii}^{\mathrm{L}} &= \mathbf{a}_{ii}-\mathbf{a}_{i,i-1}(\mathbf{d}_{i-1,i-1}^{\mathrm{L}})^{-1}\mathbf{a}_{i-1,i} & \mathbf{c}_i^{\mathrm{L}} &= -\mathbf{a}_{i+1,i}(\mathbf{d}_{ii}^{\mathrm{L}})^{-1} \\ &\vdots & &\vdots \\ \mathbf{d}_{nn}^{\mathrm{L}} &= \mathbf{a}_{nn}-\mathbf{a}_{n,n-1}(\mathbf{d}_{n-1,n-1}^{\mathrm{L}})^{-1}\mathbf{a}_{n-1,n} & \mathbf{c}_{n-1}^{\mathrm{L}} &= -\mathbf{a}_{n,n-1}(\mathbf{d}_{n-1,n-1}^{\mathrm{L}})^{-1} \end{aligned}$$

The matrices $\mathbf{d}_{ii}^{\mathrm{L}}$ are the diagonal blocks of the resulting matrix on the left. It can be seen that each diagonal block is calculated from the following relation:

$$\mathbf{d}_{ii}^{\mathrm{L}} = \mathbf{a}_{ii} + \mathbf{c}_{i-1}^{\mathrm{L}}\mathbf{a}_{i-1,i}, \quad \text{where } i = 2,3,\ldots,n \text{ and } \mathbf{d}_{11}^{\mathrm{L}} = \mathbf{a}_{11}, \tag{4}$$

and each row multiplication factor is:

$$\mathbf{c}_i^{\mathrm{L}} = -\mathbf{a}_{i+1,i}(\mathbf{d}_{ii}^{\mathrm{L}})^{-1}, \quad \text{where } i = 1,2,\ldots,n-1. \tag{5}$$

The similar backward procedure is characterized with the superscript R since the elimination procedure moves from the *right* electrode towards the left. The derivation of the backwards recursive expressions follows that of the forward elimination. Each diagonal block can be calculated from the following relation:

$$\mathbf{d}_{ii}^{\mathrm{R}} = \mathbf{a}_{ii} + \mathbf{c}_{i+1}^{\mathrm{R}}\mathbf{a}_{i+1,i}, \quad \text{where } i = n-1,\ldots,2,1 \text{ and } \mathbf{d}_{nn}^{\mathrm{R}} = \mathbf{a}_{nn}, \tag{6}$$

and each row multiplication factor is:

$$\mathbf{c}_i^{\mathrm{R}} = -\mathbf{a}_{i-1,i}(\mathbf{d}_{ii}^{\mathrm{R}})^{-1}, \quad \text{where } i = n, \ldots, 3, 2. \tag{7}$$

3. Combining the two procedures

After a complete forward and backward block Gaussian elimination sweep, the augmented matrices, named $[\mathbf{D}^{\mathrm{L}}|\mathbf{J}^{\mathrm{L}}]$ and $[\mathbf{D}^{\mathrm{R}}|\mathbf{J}^{\mathrm{R}}]$, respectively, will look as follows where the matrices $\mathbf{J}^{\mathrm{L}}$ and $\mathbf{J}^{\mathrm{R}}$ are lower and upper block triangular, respectively:

$$[\mathbf{D}^{\mathrm{L}} \mid \mathbf{J}^{\mathrm{L}}] = \left(\begin{array}{ccccc|ccc} \mathbf{d}_{11}^{\mathrm{L}} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & \\ \mathbf{0} & \mathbf{d}_{22}^{\mathrm{L}} & \mathbf{a}_{23} & & & \mathbf{c}_1^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{i}_{22} & \\ & \mathbf{0} & \mathbf{d}_{33}^{\mathrm{L}} & \mathbf{a}_{34} & & \mathbf{c}_{2,1}^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{c}_2^{\mathrm{L}}\mathbf{i}_{22} & \mathbf{i}_{33} \\ & & \ddots & \ddots & \ddots & \vdots & \vdots & \ddots \end{array}\right), \tag{8}$$

$$[\mathbf{D}^{\mathrm{R}} \mid \mathbf{J}^{\mathrm{R}}] = \left(\begin{array}{ccccc|ccc} \mathbf{d}_{11}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{11} & \mathbf{c}_2^{\mathrm{R}}\mathbf{i}_{22} & \mathbf{c}_{2,3}^{\mathrm{R}}\mathbf{i}_{33} & \cdots \\ \mathbf{a}_{21} & \mathbf{d}_{22}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{22} & \mathbf{c}_3^{\mathrm{R}}\mathbf{i}_{33} & \cdots \\ & \mathbf{a}_{32} & \mathbf{d}_{33}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{33} & \cdots \\ & & \ddots & \ddots & \ddots & & & & \ddots \end{array}\right). \tag{9}$$

Here, the following notation was introduced:

$$\left.\begin{array}{l} \mathbf{c}_1^{\mathrm{R}}\mathbf{c}_2^{\mathrm{R}}\cdots\mathbf{c}_i^{\mathrm{R}} = \mathbf{c}_{1,2,\ldots,i}^{\mathrm{R}} \\ \mathbf{c}_i^{\mathrm{L}}\mathbf{c}_{i-1}^{\mathrm{L}}\cdots\mathbf{c}_1^{\mathrm{L}} = \mathbf{c}_{i,i-1,\ldots,1}^{\mathrm{L}} \end{array}\right\} \quad \text{where } i = 1, 2, \ldots, n. \tag{10}$$

Combining the results obtained from Eqs. (2), (8), and (9) by employing the fact that

$$\mathbf{AG} = \mathbf{I}, \quad \mathbf{D}^{\mathrm{L}}\mathbf{G} = \mathbf{J}^{\mathrm{L}}, \quad \mathbf{D}^{\mathrm{R}}\mathbf{G} = \mathbf{J}^{\mathrm{R}}, \tag{11}$$

the expression

$$(\mathbf{A} - \mathbf{D}^{\mathrm{L}} - \mathbf{D}^{\mathrm{R}})\mathbf{G} = \mathbf{I} - \mathbf{J}^{\mathrm{L}} - \mathbf{J}^{\mathrm{R}} \tag{12}$$

is examined, which can be viewed as the following augmented matrix expression:

$$\left[\begin{array}{c|c} \mathbf{B} & \mathbf{F} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A} & \mathbf{I} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{D}^{\mathrm{L}} & \mathbf{J}^{\mathrm{L}} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{D}^{\mathrm{R}} & \mathbf{J}^{\mathrm{R}} \end{array}\right], \tag{13}$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{a}_{11} - \mathbf{d}_{11}^{\mathrm{L}} - \mathbf{d}_{11}^{\mathrm{R}} & & & \\ & \mathbf{a}_{22} - \mathbf{d}_{22}^{\mathrm{L}} - \mathbf{d}_{22}^{\mathrm{R}} & & \\ & & \mathbf{a}_{33} - \mathbf{d}_{33}^{\mathrm{L}} - \mathbf{d}_{33}^{\mathrm{R}} & \\ & & & \ddots \end{pmatrix} \tag{14}$$

and

$$\mathbf{F} = \begin{pmatrix} -\mathbf{i}_{11} & -\mathbf{c}_{2}^{\mathrm{R}} & -\mathbf{c}_{2,3}^{\mathrm{R}} & -\mathbf{c}_{2,3,4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{1}^{\mathrm{L}} & -\mathbf{i}_{22} & -\mathbf{c}_{3}^{\mathrm{R}} & -\mathbf{c}_{3,4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{2,1}^{\mathrm{L}} & -\mathbf{c}_{2}^{\mathrm{L}} & -\mathbf{i}_{33} & -\mathbf{c}_{4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{3,2,1}^{\mathrm{L}} & -\mathbf{c}_{3,2}^{\mathrm{L}} & -\mathbf{c}_{3}^{\mathrm{L}} & -\mathbf{i}_{44} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{15}$$

When $\mathbf{B}$ is subsequently reduced to the identity matrix $\mathbf{I}$, $\mathbf{F}$ will simultaneously be transformed into the Green's function matrix $\mathbf{G}$. In other words, the Green's function matrix sought for can be expressed as $\mathbf{G} = \mathbf{B}^{-1}\mathbf{F}$. The Green's function matrix is:

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_{11} & \mathbf{g}_{11}\mathbf{c}_{2}^{\mathrm{R}} & \mathbf{g}_{11}\mathbf{c}_{2,3}^{\mathrm{R}} & \cdots & \mathbf{g}_{11}\mathbf{c}_{2,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{22}\mathbf{c}_{1}^{\mathrm{L}} & \mathbf{g}_{22} & \mathbf{g}_{22}\mathbf{c}_{3}^{\mathrm{R}} & \cdots & \mathbf{g}_{22}\mathbf{c}_{3,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{33}\mathbf{c}_{2,1}^{\mathrm{L}} & \mathbf{g}_{33}\mathbf{c}_{2}^{\mathrm{L}} & \mathbf{g}_{33} & \cdots & \mathbf{g}_{33}\mathbf{c}_{4,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{44}\mathbf{c}_{3,2,1}^{\mathrm{L}} & \mathbf{g}_{44}\mathbf{c}_{3,2}^{\mathrm{L}} & \mathbf{g}_{44}\mathbf{c}_{3}^{\mathrm{L}} & \cdots & \mathbf{g}_{44}\mathbf{c}_{5,\ldots,n}^{\mathrm{R}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,1}^{\mathrm{L}} & \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,2}^{\mathrm{L}} & \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,3}^{\mathrm{L}} & \cdots & \mathbf{g}_{nn} \end{pmatrix}, \tag{16}$$

where the following expression for the diagonal blocks of the Green's function matrix is introduced:

$$\mathbf{g}_{ii} = -\mathbf{b}_{ii}^{-1} = \left(-\mathbf{a}_{ii} + \mathbf{d}_{ii}^{\mathrm{L}} + \mathbf{d}_{ii}^{\mathrm{R}}\right)^{-1} \quad \text{where } i = 1, 2, \ldots, n. \tag{17}$$

Off-diagonal entries are then calculated via appropriate multiplications with calculated diagonal block matrices and factors obtained during block Gaussian elimination as follows using the notation given in Eq. (10):

$$\mathbf{g}_{ij} = \mathbf{g}_{ii}\mathbf{c}_{i+1,i+2,\ldots,j-1,j}^{\mathrm{R}} \quad \text{for } i < j \tag{18}$$

$$\mathbf{g}_{ij} = \mathbf{g}_{ii}\mathbf{c}_{i-1,i-2,\ldots,j+1,j}^{\mathrm{L}} \quad \text{for } i > j. \tag{19}$$

4. Computation of transmission

The calculation of transmission t, given by the following Fisher–Lee [24] relation obtained in non–equilibrium Green's function theory, can be expressed as (cf. [21,25]):

$$t = \mathrm{Tr}\{\mathbf{G}\boldsymbol{\Gamma}^{\mathrm{L}}\mathbf{G}^{\dagger}\boldsymbol{\Gamma}^{\mathrm{R}}\}. \tag{20}$$

Here 'Tr' denotes a matrix *trace* operation, and the dagger denotes Hermitian conjugation. Regarding $\boldsymbol{\Gamma}^{\mathrm{L}}$ and $\boldsymbol{\Gamma}^{\mathrm{R}}$, the superscripts indicate *left* and *right* electrode contact leads. These matrices are defined from the electrode self-energy [21]:

$$\boldsymbol{\Gamma}^{\mathrm{L}} = \hat{\imath}\left(\boldsymbol{\Sigma}^{\mathrm{L}} - \left(\boldsymbol{\Sigma}^{\mathrm{L}}\right)^{\dagger}\right), \quad \boldsymbol{\Gamma}^{\mathrm{R}} = \hat{\imath}\left(\boldsymbol{\Sigma}^{\mathrm{R}} - \left(\boldsymbol{\Sigma}^{\mathrm{R}}\right)^{\dagger}\right), \tag{21}$$

where $\hat{\imath}$ is the imaginary unit. These matrices are only non–zero in the $(1,1)$ block for $\boldsymbol{\Sigma}^{\mathrm{L}}$ and $\boldsymbol{\Gamma}^{\mathrm{L}}$, and in the (n,n) block for the case $\boldsymbol{\Sigma}^{\mathrm{R}}$ and $\boldsymbol{\Gamma}^{\mathrm{R}}$ (cf. [26–28]).

Two methods are now presented that can be used to calculate the transmission t given in Eq. (20).

4.1. Coupling method

The coupling method is by far the popular method of choice in the literature when transmission is to be calculated via the Green's function formalism (see [26–28]). The method is introduced here, and regarded as the *baseline* method to compare the new transmission calculation method to later in the paper.

In this method, the coupling between the left and right leads is calculated, and the transmission computed accordingly. This coupling is denoted as $\mathbf{g}_{n1}$, and it resides as the lowest left corner of the Green's function matrix $\mathbf{G}$. The calculation of transmission for a particular energy ε then becomes (cf. [26]):

$$t = \mathrm{Tr}\{\mathbf{g}_{n1}\boldsymbol{\gamma}^{\mathrm{L}}_{11}\mathbf{g}^{\dagger}_{n1}\boldsymbol{\gamma}^{\mathrm{R}}_{nn}\}, \tag{22}$$

where $\boldsymbol{\gamma}^{\mathrm{L}}_{11} = [\boldsymbol{\Gamma}^{\mathrm{L}}]_{11}$ and $\boldsymbol{\gamma}^{\mathrm{R}}_{nn} = [\boldsymbol{\Gamma}^{\mathrm{R}}]_{nn}$. Thus we introduce the notation $[\cdot]_{ij}$ which delivers the (i,j)-block, with respect to $\mathbf{A}$'s block structure, of the bracketed expression. The main task is to find $\mathbf{g}_{n1}$. From Eq. (16) it is seen that the expression for this matrix is:

$$\mathbf{g}_{n1} = \mathbf{g}_{nn}\mathbf{c}^{\mathrm{L}}_{n-1}\mathbf{c}^{\mathrm{L}}_{n-2}\cdots\mathbf{c}^{\mathrm{L}}_{2}\mathbf{c}^{\mathrm{L}}_{1}, \tag{23}$$

and we see that the only factors $\mathbf{c}^{\mathrm{L}}_{i}$ involved are all those computed in a downwards block Gaussian elimination sweep. The matrix $\mathbf{g}_{nn}$ in Eq. (23) can be obtained by considering the nth block from Eq. (17):

$$\mathbf{g}_{nn} = (-\mathbf{a}_{nn} + \mathbf{d}^{\mathrm{L}}_{nn} + \mathbf{d}^{\mathrm{R}}_{nn})^{-1} = (\mathbf{d}^{\mathrm{L}}_{nn})^{-1}, \tag{24}$$

since $\mathbf{d}^{\mathrm{R}}_{nn} = \mathbf{a}_{nn}$. This holds similarly for the first row of the Green's function matrix. From this, it is seen that the first and last diagonal blocks of the Green's function matrix correspond to the final blocks of upwards and downwards sweeps of block Gaussian elimination, respectively, in the following manner:

$$\mathbf{g}_{11} = (\mathbf{d}^{\mathrm{R}}_{11})^{-1} \quad \text{and} \quad \mathbf{g}_{nn} = (\mathbf{d}^{\mathrm{L}}_{nn})^{-1}. \tag{25}$$

4.2. Overlap method

A new method that seeks to compute the transmission much like the baseline coupling method, however via a different part of the Green's function matrix, is now introduced.

Here, the idea is again based on the transmission formula Eq. (20), however the matrices dealt with change from being a coupling between the leads to that of a coupling between two adjacent blocks somewhere in the center of the system. This corresponds to centering calculations around a diagonal block of $\mathbf{A}$. This will require us to calculate the Green's function for the kth block of interest, $\mathbf{g}_{kk}$.

The name of the method arises from the fact that calculation of a diagonal block involves a *sweep* of block Gauss elimination from both the upper left and lower right of $\mathbf{A}$ which will *overlap* on the block of interest.

The motivation behind this approach is to avoid the work in having to calculate an off-diagonal block of the Green's function matrix after a series of block Gaussian elimination sweeps. This amounts to $n-1$ matrix multiplications. In the new method, overhead will arise due to calculations involving the self-energies $\boldsymbol{\Sigma}^{\mathrm{L}}$ and $\boldsymbol{\Sigma}^{\mathrm{R}}$, and the corresponding matrices $\boldsymbol{\Gamma}^{\mathrm{L}}$ and $\boldsymbol{\Gamma}^{\mathrm{R}}$. However, these computations are less expensive matrix addition operations, and they are negligible with increasing number of matrix blocks and block sizes.

As we shall demonstrate below, it is advantageous to choose k corresponding to the smallest diagonal block inside the block tridiagonal matrix $\mathbf{A}$. Although this approach involves some additional computations with the self-energy matrices and their corresponding coupling matrices, this overhead is acceptable due to the savings involved in the cheaper matrix computations for the overlap method.

Choosing an arbitrary kth diagonal block, the transmission is given in the following expression, derived in the appendix:

$$t = \mathrm{Tr}\{\mathbf{g}_{kk}[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}\mathbf{g}^{\dagger}_{kk}[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}\}, \tag{26}$$

where the new self-energy related terms are given by Eqs. (48) and (49) in the Appendix. Using the nonzero structure of the respective self-energies $\sum^{\mathrm{L}}$ and $\sum^{\mathrm{R}}$ we obtain the simpler relations

$$\left[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow 1}\right]_{11} = \boldsymbol{\gamma}^{\mathrm{L}}_{11}, \qquad \left[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow 1}\right]_{11} = \hat{\imath}((\mathbf{d}^{\mathrm{R}}_{11})^{\dagger} - \mathbf{d}^{\mathrm{R}}_{11}) \tag{27}$$

$$\left[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow n}\right]_{nn} = \boldsymbol{\gamma}^{\mathrm{R}}_{nn}, \qquad \left[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow n}\right]_{nn} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{nn})^{\dagger} - \mathbf{d}^{\mathrm{L}}_{nn}) \tag{28}$$

and for $k = 2,\ldots,n-1$

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^{\dagger} - \mathbf{d}^{\mathrm{L}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{R}}_{kk}, \quad [\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{R}}_{kk})^{\dagger} - \mathbf{d}^{\mathrm{R}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{L}}_{kk}. \tag{29}$$

5. Benchmark results

The methods introduced here were implemented in C++ within Atomistix's Atomistix ToolKit, and computing times were obtained for calculating the transmission for 10 different energies ε for several different systems. These systems have been taken from the literature, and an overview of selected examples is presented in Table 1.

5.1. Operation count

In order to determine which transmission method may be algorithmically more efficient, the quantity of matrix factorizations, multiplications and additions related to the three different methods available is recorded in Table 2.

In Table 3 operation counts for the calculations of the full inverse of **A** as well as calculation of only the block tridiagonal part of the inverse is included. This is done for both a Gauss elimination (GE) algorithm, as well as the new method presented in this paper.

The block tridiagonal part of the inverse is of interest for further calculations carried out in Density Functional Theory (DFT) via the Green's function formalism, and results for the full inverse are included in order to show how the new method in this paper, though suited for the block tridiagonal calculation, is ill-suited to calculate the entire inverse, compared to traditional methods.

Looking at operation counts in Table 2 on obtaining various parts of the Green's function matrix **G**, it is seen that all choices require n LU factorizations, where n is the number of diagonal blocks in **A**.

Table 1
An overview of the test examples examined in this paper

Example systems				
System	Article	Order	n	Block order
`All00+C7`	[15]	444	5	128, 72, 16, 100, 128
`AlLead+C7`	[15]	296	5	72, 72, 20, 60, 72
`Aulll-AR`	[17]	1295	10	243, 162, 66, 79, 69, 84, 62, 62, 225, 243
`Aulll-TW`	[16]	1155	8	243, 162, 62, 70, 53, 70, 252, 243
`Aulll-DTB`	[18]	928	5	243, 162, 88, 198, 243
`Fe-MgO-Fe`	[13,14]	228	5	54, 45, 30, 45, 54
`nanotube4_4`	–	576	4	128, 128, 192, 128

For each example the original paper related to the system, the dimension of the overall matrix **A**, the number of diagonal blocks n, and finally the size of each of the diagonal blocks, from the upper left of **A** down to the lower right is listed.

Table 2
This table illustrates the amount of basic operations performed in calculating different blocks of **G** via either block Gauss elimination (GE), the coupling method or overlap method

Green's function sub-block operation count				
Block	Method	LU-factorizations	Multiplications	Additions
$\mathbf{g}_{n1}$	GE	n	$3(n-1)$	$n-1$
$\mathbf{g}_{nn}$	GE	n	$2n-1$	$n-1$
$\mathbf{g}_{kk}$	GE	n	$4n-2k-1$	$2n-k-1$
$\mathbf{g}_{n1}$	Coupling	n	$3(n-1)$	$n-1$
$\mathbf{g}_{nn}$	Overlap	n	$2n-1$	$n-1$
$\mathbf{g}_{kk}$	Overlap	n	$2n-1$	$n+1$

The third, fourth and fifth columns refer to the basic matrix operations of LU-factorization, multiplication and addition. The term n is the total amount of diagonal blocks in **A**, and k indicates which diagonal block in the Green's function matrix **G** is used for transmission calculations.

Table 3
This table illustrates the amount of basic operations performed in calculating either the full inverse **G** of **A**, or only the block tridiagonal part of it, using different methods

Green's function operation count				
Calculation	Method	LU-factorizations	Multiplications	Additions
Full inv	GE	n	$2n^2 + n - 2$	$\frac{1}{2}(n^2 + n - 2)$
Trid inv	GE	n	$\frac{1}{2}(3n^2 + 5n - 6)$	$\frac{1}{2}(n^2 + n - 2)$
Full inv	Paper	$3n - 2$	$n^2 + 4n - 4$	$4n - 6$
Trid inv	Paper	$3n - 2$	$7n - 6$	$4n - 6$

The methods employed are block Gauss Elimination (*GE*), and the new method incorporating forward and backward Gaussian elimination sweeps (*paper*), as presented in Eq. (16). The third, fourth and fifth columns refer to the basic matrix operations of LU–factorization, multiplication and addition. The term *full inv* refers to calculating the full inverse, while *trid inv* refers to obtaining only the block tridiagonal part of the inverse. The term n is the total amount of diagonal blocks in **A**.

In obtaining $\mathbf{g}_{n1}$, block Gauss elimination and the coupling method both require the same amount of operations to complete, and there is no advantage either way. Again, in obtaining the lower diagonal block $\mathbf{g}_{nn}$, both block Gauss elimination and the overlap method require the same amount of matrix–matrix calculations.

The advantage of the overlap method over block Gauss elimination occurs when a central diagonal block $\mathbf{g}_{kk}$ is required. Here, only two more matrix–matrix additions over the overlap method for $\mathbf{g}_{nn}$ is needed, while for block Gauss elimination, a series of matrix–matrix multiplies and additions add up in order to back–solve up towards the desired diagonal block. Thus the overlap method is better suited for determining diagonal blocks than block Gauss elimination.

Looking at which block of the matrix **G** is cheapest to compute on the basis of Table 2, one would apparently choose $\mathbf{g}_{nn}$. This, however, may not be the case since the table does not take into account differing block sizes among the different sub-blocks in **A** and **G**. These differing sizes can lead to substantial changes in costs regarding the basic operations of LU-factorization, matrix multiplication and matrix addition in the table. The speedup results presented later in Section 5.2 and Table 4 will verify this.

With regard to the cost of the basic operations on a matrix block of order n_i, then the amount of work for each LU-factorization, multiplication and addition is on the order of $2/3n_i^3$, $2n_i^3$ and $2n_i^2$, respectively.

5.1.1. Transmission calculation

To finally calculate transmission after successfully obtaining a sub-block of **G**, the Fisher–Lee relation (cf. Eq. (20)) is invoked, and thus three matrix–matrix multiplications are incurred, as well as a matrix trace operation. However, the significant factor here among the different methods reviewed is that the final matrix block dimensions in the Fisher–Lee relation may be different. Typically, due to the topology of the two-probe system, the central region, and thus the kth diagonal block $\mathbf{g}_{kk}$, will be of smaller size than the corner blocks $\mathbf{g}_{nn}$ or $\mathbf{g}_{n1}$. Thus a significant prefactor cost in execution time can be saved by selecting the transmission method centered around the smallest Green's function diagonal matrix block.

Table 4
This table illustrates the speedup achieved by using the new methods centered around diagonal blocks, relative to the baseline coupling method using the off-diagonal block $\mathbf{g}_{n1}$

Speedup measurements			
System	Coupling – $\mathbf{g}_{n1}$	Overlap – $\mathbf{g}_{nn}$	Overlap – $\mathbf{g}_{kk}$
`AllOO+C7`	1.0000	1.2099	2.6557
`AlLead+C7`	1.0000	1.1916	2.0092
`Aulll–AR`	1.0000	1.4211	3.2121
`Aulll–TW`	1.0000	1.3721	2.8994
`Aulll–DTB`	1.0000	1.3675	3.2537
`Fe–MgO–Fe`	1.0000	1.3064	1.8001
`nanotube4_4`	1.0000	1.2261	1.2477

The expression $\mathbf{g}_{n1}$ refers to the coupling method, while $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$ refer to the overlap method performed on the nth and smallest diagonal block, respectively. The overlap methods are always faster, and in particular those centred on the smallest, kth, diagonal block.

Some overhead arises in choosing a central diagonal block in the shape of recalculating new matrices $[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}$ and $[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}$ for the transmission function Eq. (26) via Eqs. (27)–(29), but as these operations are cheaper matrix–matrix addition operations on small matrices, this overhead is offset by the gains in being able to employ smaller matrices in the more expensive matrix–matrix multiplication operations in the Fisher–Lee relation in Eq. (26).

5.1.2. Full inversion

With regard to determining the full inverse **G** from **A**, it is seen in Table 3 how block Gauss elimination excels over the method in this paper in terms of costly LU factorizations. Although Gauss elimination has about twice the number of matrix multiplies than the new method, Gauss elimination is still preferable when taking into account that it only requires about a third LU factorizations compared to the new method. Thus the new method is not suited for determining the full matrix **G**.

However, when requiring only the tridiagonal part of the inverse, as is the case for some DFT applications, the new method is a better choice since it only requires on the order of n matrix–matrix multiplications, while block Gauss elimination still requires on the order of n^2 matrix–matrix multiplications.

5.2. Speedup results

For an overview of the speedup of the new methods relative to the baseline coupling method, see Table 4. Overall, speedup improves in every case when moving from the coupling method to the overlap method. This is not surprising, seeing how the main difference between these two methods, operation count–wise, is the lack of extra matrix multiplications in order to obtain an off-diagonal Green's function matrix block. Eliminating this task will always lead to a faster method.

Performing calculations using the smallest diagonal block k over the first or nth block can also yield significant improvements in execution time, depending on the topology of the two-probe system, and the subsequent block structure in **A**. The difference here is that it is no longer possible to 'recycle' one of the self-energy terms that is assumed to be available from the outset, as well as different block size between $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$. Thus in seeking a smaller diagonal matrix block to work with, appropriate self-energy terms must be determined once again, and this leads to extra overhead.

However, it may pay off to select some central diagonal block over a corner diagonal block in order to calculate transmission. This comes in the form of being able to work with smaller matrices, and thus matrix operation costs decrease. Crucially, matrix sizes may decrease such that memory requirements for matrix operations can be fulfilled by lower level hardware caches, leading to significant speedup in execution time. This effect is visible in Table 5, where significant speedup is achieved in the matrix–matrix operations involved in the Fisher–Lee calculation.

Table 5
This table illustrates the speedup in the calculation of solely the Fisher–Lee relation (see Eq. (26)) achieved by using the new methods centered around diagonal blocks, relative to the baseline coupling method using the off-diagonal block $\mathbf{g}_{n1}$

Speedup measurements – Fisher–Lee				
System	Coupling – $\mathbf{g}_{n1}$	Overlap – $\mathbf{g}_{nn}$	Overlap – $\mathbf{g}_{kk}$	Theoretical – $\frac{n^3}{m^3}$
`Al100+C7`	1.0000	1.0567	548.4500	512.000
`AlLead+C7`	1.0000	0.9546	47.4516	46.656
`Aulll-AR`	1.0000	1.2654	170.6912	60.207
`Aulll-TW`	1.0000	1.2788	275.6198	96.381
`Aulll-DTB`	1.0000	1.2716	59.8502	21.056
`Fe-MgO-Fe`	1.0000	1.3186	7.1354	5.832
`nanotube4_4`	1.0000	0.9940	1.0178	1.000

The expression $\mathbf{g}_{n1}$ refers to the coupling method, while $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$ refer to the overlap method performed on the nth and smallest diagonal block, respectively. The final column indicates the theoretical speedup based on the $\mathcal{O}(n^3)$ cost of evaluating Eq. (26). The reason for better speedup over theoretical prediction is due to improved cache usage by the smaller matrices dealt with when using $\mathbf{g}_{kk}$.

Furthermore, as will be explored in Section 6 concerning transmission accuracy, depending on the system, central matrix blocks may be less prone to perturbation from inaccurately calculated electrode surface Green's function matrices. This, however, varies from system to system, as well as incoming electron wave energies ε.

6. Transmission accuracy

It has been shown that any block of the Green's function matrix can be used in order to calculate transmission and a new strategy employing diagonal blocks of $\mathbf{G}$ was developed. The question now is which part of $\mathbf{G}$ might be used in order to achieve best accuracy in determining transmission. This section suggests that an investigation of the accuracy achieved for a given block may lead to informed choices. The problem of the selection of which matrix block is best concerning accuracy comes from the fact that in practice the self-energies of the electrodes, $\boldsymbol{\sigma}_{11}^{\mathrm{L}}$ and $\boldsymbol{\sigma}_{nn}^{\mathrm{R}}$, are not computed exactly. This is because in the Green's function formalism approach, the surface Green's function matrices for the electrodes (and hence their corresponding self-energies) are typically determined through an iterative procedure [29] that only converges to the correct retarded Green's function matrix when a small positive imaginary perturbation is applied. This means that transmissions are calculated for a slightly perturbed matrix $\widetilde{\mathbf{A}}$, where the corner blocks $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$ are perturbed to some degree through the inexact self-energies.

The matrix $\mathbf{A}$ here will denote the case when no imaginary perturbation is used and this can be done by employing a different manner to converge the surface Green's function matrices, such as a wave function matching [30–32] approach. To investigate how this imaginary perturbation ultimately affects the Green's function matrix that transmissions are calculated with, the inverses of an unperturbed case and a perturbed case are compared. The perturbation on $\mathbf{A}$ is described as the added matrix $\mathbf{P}$, defined as zero everywhere, except the corner blocks $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$, that correspond to the corner blocks $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$, both in size and location.

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{P}, \quad \text{where} \quad \mathbf{P} = \begin{pmatrix} \mathbf{p}_{11} & & \\ & & \\ & & \mathbf{p}_{nn} \end{pmatrix} \tag{30}$$

The perturbation matrix, as seen in Eq. (30), is divided into 9 blocks, where the empty space denotes areas with elements equal to zero. In a similar manner, the inverse $\mathbf{G} = \mathbf{A}^{-1}$ is subdivided into the same block sizes.

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & & \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{a}_{23} & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} \\ & & \ddots & \ddots & \ddots \\ & & & & \mathbf{a}_{n-1,n} \\ & & & \mathbf{a}_{n,n-1} & \mathbf{a}_{nn} \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} \mathbf{g}_{11} & \bullet & \bullet & \bullet & \mathbf{g}_{1n} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \mathbf{g}_{n1} & \bullet & \bullet & \bullet & \mathbf{g}_{nn} \end{pmatrix} \tag{31}$$

To investigate the effect of the perturbation $\mathbf{P}$ the derivation of $\widetilde{\mathbf{G}} = \widetilde{\mathbf{A}}^{-1}$ is carried out:

$$\widetilde{\mathbf{G}} = [\mathbf{A}(\mathbf{I} + \mathbf{G}\mathbf{P})]^{-1} = (\mathbf{I} + \mathbf{G}\mathbf{P})^{-1}\mathbf{G}. \tag{32}$$

If the perturbation is assumed to be small, such that the spectral radius satisfies $\rho(\mathbf{G}\mathbf{P}) < 1$, then the first inverse term can be expressed via a geometric series.

$$\widetilde{\mathbf{G}} = (\mathbf{I} + \mathbf{GP})^{-1}\mathbf{G} = \mathbf{G} - \mathbf{GPG} + \mathbf{GPGPG} - \cdots \tag{33}$$

Thus it can be seen that the difference in the perturbed and unperturbed inverses should be dominated by the term **GPG**. If **G** is subdivided into row and column blocks, as follows, it will be possible to proceed and derive a relatively compact expression for the structure of this first order correction term.

$$\mathbf{G} = \left(\begin{array}{c|c|c} \mathbf{g}_{11} & \cdots & \mathbf{g}_{1n} \\ \hline \vdots & \ddots & \vdots \\ \hline \mathbf{g}_{n1} & \cdots & \mathbf{g}_{nn} \end{array}\right) = \left(\begin{array}{c|c|c} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{array}\right) = \left(\begin{array}{c} \mathbf{c}_1 \\ \hline \vdots \\ \hline \mathbf{c}_n \end{array}\right) \tag{34}$$

such that

$$\mathbf{b}_1 = \begin{pmatrix} \mathbf{g}_{11} \\ \vdots \\ \mathbf{g}_{n1} \end{pmatrix}, \quad \mathbf{b}_n = \begin{pmatrix} \mathbf{g}_{n1} \\ \vdots \\ \mathbf{g}_{nn} \end{pmatrix} \tag{35}$$

and

$$\mathbf{c}_1 = \begin{pmatrix} \mathbf{g}_{11} & \cdots & \mathbf{g}_{1n} \end{pmatrix}, \quad \mathbf{c}_n = \begin{pmatrix} \mathbf{g}_{n1} & \cdots & \mathbf{g}_{nn} \end{pmatrix}. \tag{36}$$

With this notation, the first order perturbation term is written as follows:

$$\mathbf{GPG} = \mathbf{b}_1\mathbf{p}_{11}\mathbf{c}_1 + \mathbf{b}_n\mathbf{p}_{nn}\mathbf{c}_n. \tag{37}$$

It can be seen how the outer-product form of this expression will yield a dense matrix **GPG**, since **G** can generally be assumed to be dense. This indicates that the correction term's effect will depend directly on the full structure of **G**, and thus no prediction can be made about the effect of the perturbation on **G**, without calculating **G** itself.

We look at the first order perturbation at block (i, j):

$$[\mathbf{GPG}]_{ij} = [\mathbf{b}_1\mathbf{p}_{11}\mathbf{c}_1 + \mathbf{b}_n\mathbf{p}_{nn}\mathbf{c}_n]_{ij} = \mathbf{g}_{i1}\mathbf{p}_{11}\mathbf{g}_{1j} + \mathbf{g}_{in}\mathbf{p}_{nn}\mathbf{g}_{nj},$$

where the element $\mathbf{g}_{in}$ describes the amplitude of an electron propagating from site i to site n in the system. For most systems, this will decay as a function of the distance between orbitals at sites i and n, and thus the error should be smallest for Green's function blocks in the center of the cell, i.e., as far as possible from the electrodes. Thus we can expect choosing central blocks in **G** should lead to more accurate transmission calculations for most systems.

6.1. Numerical example with random perturbation

The effect of a perturbation of the electrode's surface Green's function matrices on the Green's function **G** itself is here illustrated by a numerical example. The Hamiltonian matrix **H** and overlap matrix **S** associated with Au111–AR is taken, and the matrix to be inverted is constructed as

$$\mathbf{A} = \mathbf{H} - \varepsilon\mathbf{S}, \quad \text{where } \varepsilon = 1.0. \tag{38}$$

The corner blocks of **A**, namely $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$, are then perturbed with matrices $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$. The elements of $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$ are computed as:

$$\mathbf{p}_{11} \leftarrow p_{ij} = \alpha_{ij}a_{ij}, \quad \text{where } a_{ij} \in \mathbf{a}_{11}, \text{ and} \tag{39}$$

$$\mathbf{p}_{nn} \leftarrow p_{kl} = \alpha_{kl}a_{kl}, \quad \text{where } a_{kl} \in \mathbf{a}_{nn}. \tag{40}$$

where the factors α_{ij} and α_{kl} are normally distributed with zero mean and standard deviation 10^{-5}.

Fig. 3 shows the results of the average difference of 100 perturbed inversions $\widetilde{\mathbf{G}}$ compared to **G**. From this figure, it is seen that for this particular choice of system (**H** and **S**) and energy (ε), the perturbation from the iterated self-energies cause the inverse to be most inaccurate at the corner diagonal blocks. Thus, choosing the

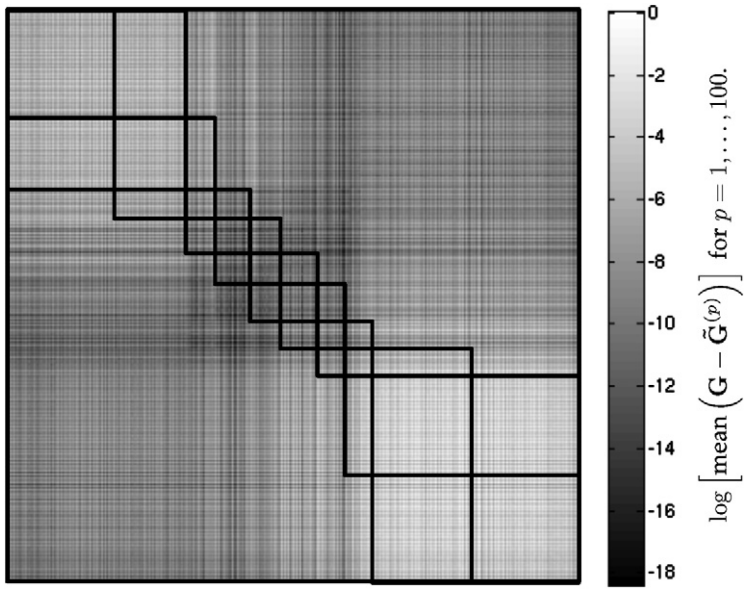

Fig. 3. The figure above illustrates the average element-wise difference expressed as $\log[\text{mean}(\mathbf{G}-\widetilde{\mathbf{G}}^{(p)})]$ for $p=1,\ldots,100$. The matrix $\mathbf{G}$ corresponds to the Au111–AR example [17]. Element-wise differences range from about the same order down to about 18 orders of magnitude smaller. The dark lines outline the original block tridiagonal structure of the original matrix $\mathbf{A}$. The logarithm employed is the base 10 logarithm. In this particular example for choice of electron energy ε and $\mathbf{A}$, the diagonal blocks in the center of the matrix suffer least in terms of accuracy.

overlap method as the transmission calculation method would be on average best served by choosing a block towards the center of the matrix, where the perturbation has the least effect. This choice is further motivated by the fact that the center blocks typically are of smaller size, and matrix operations would be faster than operations with the corner diagonal blocks.

A problem with this analogy lies in the fact that one can not predict which Green's function matrix block would provide more accurate transmission results (see Eq. (37)), without calculating the Green's function matrix in the first place. This lends prediction to be prohibitive in general, when computing transmissions. The best choice of action is then relying on the usual behavior of most two-probe systems as well as choosing the fastest calculation method, leading us to pick a diagonal block towards the center of the system, which are typically the least affected by the electrodes as well as the smallest in size.

7. Conclusion

This paper developed and introduced a new, faster method of calculating transmission for two-probe systems by using diagonal block matrices from the Green's function matrix, $\mathbf{g}_{ii}$, rather than the coupling method found extensively in the literature that uses the corner off-diagonal block $\mathbf{g}_{n1}$.

This is done by developing a method for calculating any block matrix from the Green's function matrix $\mathbf{G}$ based on a series of Gauss eliminations carried out on the original matrix $\mathbf{A}$.

To calculate transmission via a diagonal block of the Green's function matrix $\mathbf{G}$, upwards and downwards block Gaussian elimination is performed that terminates overlapping over $\mathbf{a}_{kk}$, and $\mathbf{g}_{kk}$ is calculated (cf. Eq. (17)).

Furthermore, the related coupling matrices (usually obtained via self-energy) used in the transmission formula Eq. (26) are calculated via Eqs. (27)–(29), for the new, extended electrodes. This approach dispenses with the need of a series of matrix–matrix multiplications compared to the coupling method (cf. Eq. (23)) in exchange for cheaper matrix–matrix addition operations.

Execution time measurements indicated that centering transmission calculations on the Green's function matrix's diagonal blocks was preferable, in that a series of matrix–matrix multiplications would be saved as

well as centering on smaller diagonal matrix blocks offset the cost of re-calculating self-energy matrices. Furthermore, the ability to choose smaller block matrices lends itself to the possibility of far better cache usage, and hence greater performance gains.

Perturbation analysis revealed that it is not possible to determine the effect of perturbation in the electrode self-energy matrices on the accuracy of the Green's function, without explicitly calculating the Green's function matrix. This eliminates the ability to predict which Green's function matrix block would be an ideal choice for the calculation of a two-probe system's transmission with respect to accuracy. However, due to the behavior of most two-probe systems, a central diagonal block choice is expected to yield more accurate results.

Acknowledgments

This work was supported by Grant No. 2106-04-0017, "Parallel Algorithms for Computational Nano-Science", under the NABIIT program from the Danish Council for Strategic Research.

Appendix. Derivation of Eq. (26) for the Transmission

We commence with the expression in Eq. (1). As shown in (e.g., Golub and Van Loan [33]) Section 3.2.1, we can represent a Gauss-elimination step as a matrix multiplication with a "Gauss transformation". The same is true for the block Gauss-elimination steps we use here, and thus we express a series of downwards Gauss-eliminations that terminate on row k by $\mathbf{E}_{\downarrow k}$. Similarly, a series of upwards Gauss-eliminations terminating on row k is denoted by $\mathbf{E}_{\uparrow k}$. We then write the combination of Gauss-elimination sweeps that produce a matrix $\mathbf{Z}_k$ as follows:

$$\mathbf{Z}_k = \mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}_{\uparrow k}. \tag{41}$$

Due to the structure of $\mathbf{A}$, the matrix $\mathbf{Z}_k$ is block diagonal as shown in Fig. 4. Given $\mathbf{Z}_k$, we can write the Green's function matrix as

$$\mathbf{G} = \mathbf{A}^{-1} = \mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}. \tag{42}$$

We can then insert this expression into the Fisher–Lee relation from Eq. (20), to obtain

$$\begin{aligned} t &= \mathrm{Tr}\{(\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k})\mathbf{\Gamma}^{\mathrm{L}}(\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k})^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\} = \mathrm{Tr}\{\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\} \\ &= \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}_{\uparrow k}\} = \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{\Gamma}_{\downarrow k}^{\mathrm{L}}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{\Gamma}_{\uparrow k}^{\mathrm{R}}\} \end{aligned} \tag{43}$$

where we have introduced

$$\mathbf{\Gamma}_{\downarrow k}^{\mathrm{L}} = \mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger} \quad \text{and} \quad \mathbf{\Gamma}_{\uparrow k}^{\mathrm{R}} = \mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}_{\uparrow k}. \tag{44}$$

To derive Eq. (43) we used that the trace is invariant under matrix commutation [34].

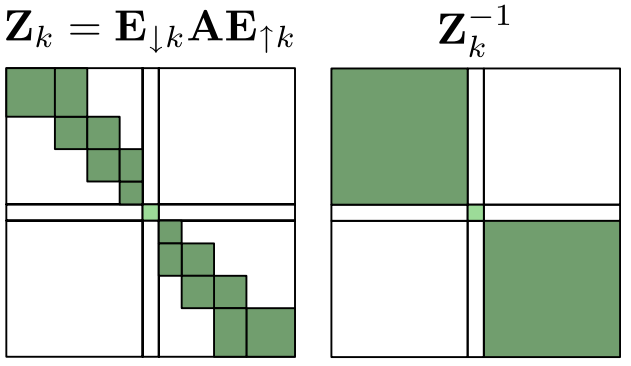

Fig. 4. The zero/nonzero structure of $\mathbf{Z}_k$ and $\mathbf{Z}_k^{-1}$.

Eq. (43) can be further simplified. First note that both $\mathbf{Z}_k$ and $\mathbf{Z}_k^{-1}$ have the special zero/nonzero structure shown in Fig. 4. Next, note that $\mathbf{\Gamma}^{\mathrm{L}}$ has nonzero elements in its (1,1)-block only, and hence the nonzeros in $\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}$ are confined to upper left blocks, as shown in Fig. 5. Similarly, the nonzeros of $\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}$ are confined to the bottom right blocks. Using the zero/nonzero structure of these matrices, it follows from the derivation illustrated in Fig. 6 that:

$$t = \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}\} = \mathrm{Tr}\{[\mathbf{Z}_k^{-1}]_{kk}[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}[\mathbf{Z}_k^{-1}]^{\dagger}_{kk}[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}\}. \tag{45}$$

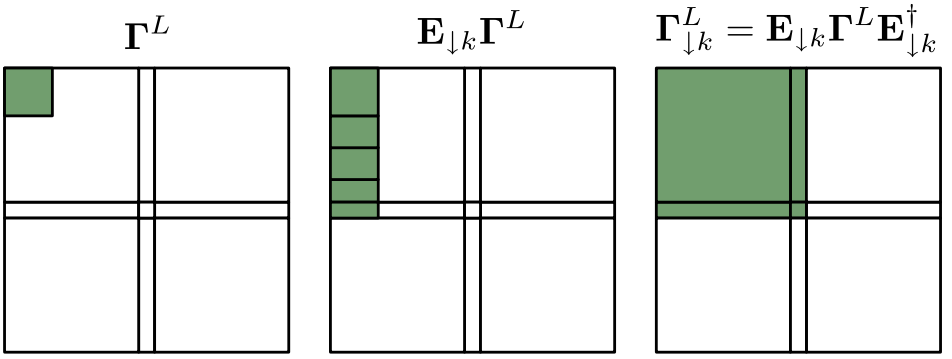

Fig. 5. The zero/nonzero structure of $\mathbf{\Gamma}^{\mathrm{L}}$ and $\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}$.

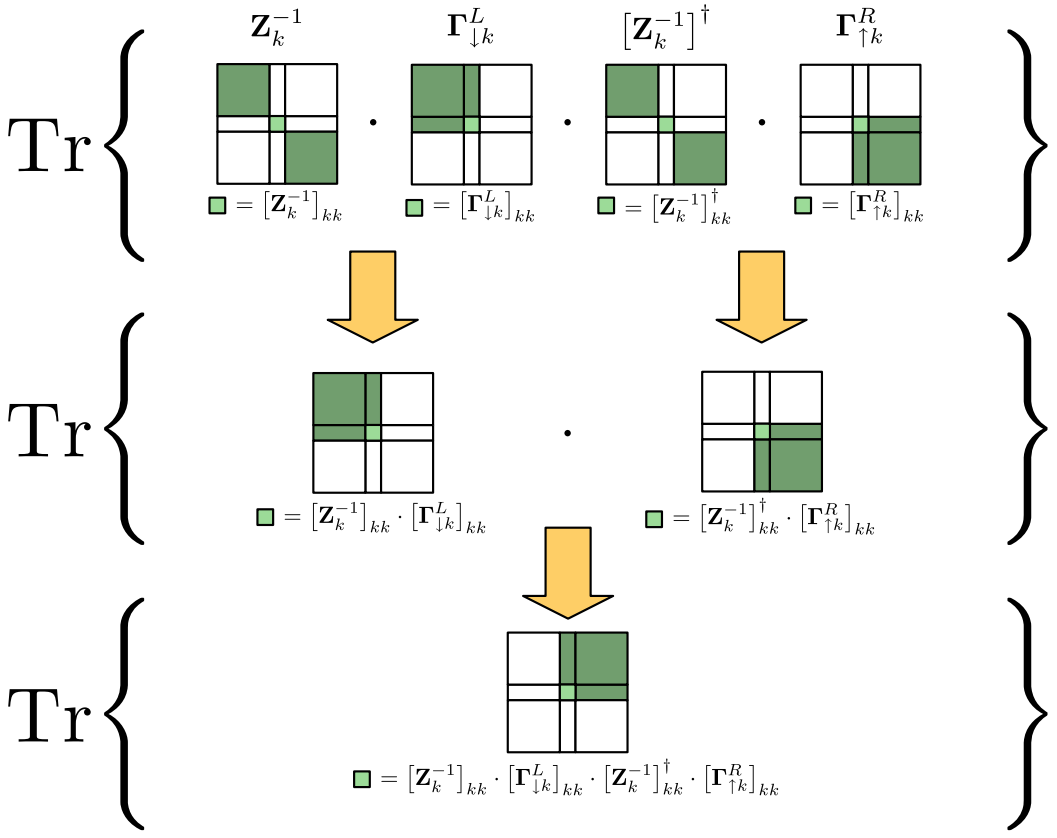

Fig. 6. Illustration of the derivation of Eq. (45) using the zero/nonzero structure of Figs. 4 and 5.

Hence, we require only the kth diagonal block of $\mathbf{Z}_k^{-1}$, and we note that this corresponds to the kth diagonal block of the Green's function matrix $\mathbf{G}$ via Eq. (17). Thus $[\mathbf{Z}_k^{-1}]_{kk} = \mathbf{g}_{kk}$, and $[\mathbf{Z}_k^{-1}]_{kk}^\dagger = \mathbf{g}_{kk}^\dagger$.

Next we consider $[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}$ and $[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}$. By means of Eq. (1) we can obtain the expression of a self-energy, e.g., $\mathbf{\Sigma}^{\mathrm{L}}$, and via Eq. (21) we now determine our desired matrix for the transmission calculation:

$$\begin{aligned}
[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} &= [\mathbf{E}_{\downarrow k}\hat{\imath}(\mathbf{\Sigma}^{\mathrm{L}} - (\mathbf{\Sigma}^{\mathrm{L}})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk} = \hat{\imath}[\mathbf{E}_{\downarrow k}((\varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{R}} - \mathbf{A}) - (\varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{R}} - \mathbf{A})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk} \\
&= \hat{\imath}[\mathbf{E}_{\downarrow k}(\mathbf{A}^\dagger - \mathbf{A} - (\mathbf{\Sigma}^{\mathrm{R}} - (\mathbf{\Sigma}^{\mathrm{R}})^\dagger))\mathbf{E}^\dagger_{\downarrow k}]_{kk} \\
&= \hat{\imath}([\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]^\dagger_{kk} - [\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]_{kk}) - \hat{\imath}[\mathbf{E}_{\downarrow k}(\mathbf{\Sigma}^{\mathrm{R}} - (\mathbf{\Sigma}^{\mathrm{R}})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk}.
\end{aligned} \tag{46}$$

Here we used that both $\mathbf{S}$ and $\mathbf{H}$ are Hermitian and therefore vanish in the expression. The first term involving $\mathbf{A}$ is simplified via the fact that the (k,k)-subblock of the block tridiagonal $\mathbf{E}_{\downarrow k}\mathbf{A}$ remains invariant under the column operations by $\mathbf{E}^\dagger_{\downarrow k}$, and thus $[\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]_{kk} = \mathbf{d}^{\mathrm{L}}_{kk}$. The last term, involving self-energies, is simplified via Eq. (21). We get

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^\dagger - \mathbf{d}^{\mathrm{L}}_{kk}) - [\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}]_{kk}. \tag{47}$$

Since $\mathbf{E}_{\downarrow k}$ represents downwards elimination, the (k,k)-block in $\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}$ is left unaltered, i.e., $[\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}]_{kk} = [\mathbf{\Gamma}^{\mathrm{R}}]_{kk} = \boldsymbol{\gamma}^{\mathrm{R}}_{kk}$. Hence:

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^\dagger - \mathbf{d}^{\mathrm{L}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{R}}_{kk}. \tag{48}$$

Following a similar procedure, we obtain:

$$[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{R}}_{kk})^\dagger - \mathbf{d}^{\mathrm{R}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{L}}_{kk}. \tag{49}$$

Thus we have all the terms necessary for the calculation of transmission via Eq. (43).

References

[1] P. Pernas, A. Martin-Rodero, F. Flores, Electrochemical–potential variations across a constriction, Phys. Rev. B 41 (1990) 8553–8556.
[2] W. Tian, S. Datta, Aharonov–Bohm-type effect in graphene tubules: a Landauer approach, Phys. Rev. B 49 (1994) 5097–5100.
[3] L. Chico, M. Sancho, M. Munoz, Carbon-nanotube-based quantum dot, Phys. Rev. Lett. 81 (1998) 1278–1281.
[4] A. de Parga, O.S. Hernan, R. Miranda, A.L. Yeyati, A. Martin-Rodero, F. Flores, Electron resonances in sharp tips and their role in tunneling spectroscopy, Phys. Rev. Lett. 80 (1998) 357–360.
[5] N.D. Lang, Resistance of atomic wires, Phys. Rev. B 52 (1995) 5335–5342.
[6] K. Hirose, M. Tsukada, First-principles calculation of the electronic structure for a bielectrode junction system under strong field and current, Phys. Rev. B 51 (1995) 5278–5290.
[7] M.B. Nardelli, Electronic transport in extended systems: application to carbon nanotubes, Phys. Rev. B 60 (1999) 7828–7833.
[8] M.B. Nardelli, J. Bernholc, Mechanical deformations coherent transport in carbon nanotubes, Phys. Rev. B 60 (1999) R16338–R16341.
[9] J.J. Palacios, A.J. Pérez-Jiménez, E. Louis, J.A. Vergés, Fullerene-based molecular nanobridges: a first-principles study, Phys. Rev. B 64 (2001) 115411.
[10] P.A. Derosa, J.M. Seminario, Electron transport through single molecules: scattering treatment using density functional and Green Function theories, J. Phys. Chem. B 105 (2001) 471–481.
[11] S.N. Yaliraki, A.E. Roitberg, C. Gonzalez, V. Mujica, M.A. Ratner, The injecting energy at molecule/metal interfaces: implications for conductance of molecular junctions from an ab initio molecular description, J. Chem. Phys. 111 (1999) 6997–7002.
[12] J. Taylor, H. Guo, J. Wang, Ab initio modeling of open systems: charge transfer, electron conduction, and molecular switching of a C_{60} device, Phys. Rev. B 63 (2001) 121104.
[13] M. Stilling, K. Stokbro, K. Flensberg, Electronic transport in crystalline magnetotunnel junctions: effects of structural disorder, J. Comput.-Aid. Mater. Des. 14 (2007) 141–149.
[14] M. Stilling, K. Stokbro, K. Flensberg, Crystalline magnetotunnel junctions: Fe–MgO–Fe, Fe–FeOMgO–Fe and Fe–AuMgOAu–Fe, Nanotech 3 (2006) 39–42.
[15] M. Brandbyge, J.L. Mozos, P. Ordejón, J. Taylor, K. Stokbro, Density-functional method for nonequilibrium electron transport, Phys. Rev. B 65 (2002) 165401.
[16] J. Taylor, M. Brandbyge, K. Stokbro, Theory of rectification in tour wires: the role of electrode coupling, Phys. Rev. Lett. 89 (2002) 138301.
[17] K. Stokbro, J. Taylor, M. Brandbyge, Do Aviram–Ratner diodes rectify? J. Amer. Chem. Soc. 125 (2003) 3674–3675.

[18] K. Stokbro, J.L. Mozos, P. Ordejón, M. Brandbyge, J. Taylor, Theoretical study of the nonlinear conductance of di-thiol benzene coupled to Au(1 1 1) surfaces via thiol and thiolate bonds, Comput. Mater. Sci. 27 (2003) 151–160.
[19] R. Hoffmann, An extended Hückel theory. I. Hydrocarbons, J. Chem. Phys. 39 (1963) 1397–1412.
[20] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (1965) A1133–A1138.
[21] S. Datta, Electronic Transport in Mesoscopic Systems, Cambridge Univ. Press, New York, 1996.
[22] E.M. Godfrin, A method to compute the inverse of an *n*-block tridiagonal quasi-Hermitian matrix, J. Phys.: Condens. Matter 3 (1991) 7843–7848.
[23] J.D. Gilbert, L. Gilbert, Linear Algebra and Matrix Theory, Academic Press Inc., 1995.
[24] D.S. Fisher, P.A. Lee, relation between conductivity and transmission matrix, Phys. Rev. B 23 (1981) 6851–6854.
[25] H. Haug, A.-P. Jauho, Quantum Kinetics in Transport and Optics of Semiconductors, Springer-Verlag, Berlin, Heidelberg, 1996.
[26] P.S. Drouvelis, P. Schmelcher, P. Bastian, Parallel implementation of the recursive Green's function method, J. Comp. Phys. 215 (2006) 741–756.
[27] S.V. Faleev, F. Léonard, D.A. Stewart, M. van Schilfgaarde, Ab initio tight-binding LMTO method for nonequilibrium electron transport in nanosystems, Phys. Rev. B 71 (2005) 195422.
[28] O. Hod, J.E. Peralta, G.E. Scuseria, First-principles electronic transport calculations in finite elongated systems: a divide and conquer approach, J. Chem. Phys. 125 (2006) 114704.
[29] M.P. López Sancho, J.M. López Sancho, J. Rubio, Highly convergent schemes for the calculation of bulk and surface Green functions, J. Phys. F: Met. Phys. 15 (1985) 851–858.
[30] H.H. Sørensen, D.E. Petersen, P.C. Hansen, S. Skelboe, K. Stokbro, Efficient wave function matching approach for quantum transport calculations, Phys. Rev. B, submitted for publication.
[31] P.A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, P.J. Kelly, Conductance calculations for quantum wires and interfaces: mode matching and Green's functions, Phys. Rev. B 72 (2005) 035450.
[32] T. Ando, Quantum point contacts in magnetic fields, Phys. Rev. B 44 (1991) 8017–8027.
[33] G.H. Golub, C.F. van Loan, Matrix Computations, Johns Hopkins University Press, London, 1996.
[34] S. Lang, Linear Algebra, Springer-Verlag, New York, 1987.