Доклады Национальной академии наук Беларуси

2015 **январь–февраль** **Том 59 № 1**

УДК 539.21

В. Е. ГУСАКОВ

РАСЧЕТ ШИРИНЫ ЗАПРЕЩЕННОЙ ЗОНЫ ПОЛУПРОВОДНИКОВ В РАМКАХ МЕТОДА ФУНКЦИОНАЛА ПЛОТНОСТИ

(Представлено академиком Н. М. Олехновичем)

НПЦ НАН Беларуси по материаловедению, Минск *Поступило 22.12.2014*

Введение. Метод теории функционала плотности (DFT) является одним из наиболее широко используемых методов в теоретической физике конденсированного состояния. Однако при расчете электронных свойств конденсированных сред в рамках метода DFT существует так называемая проблема ширины запрещенной зоны (см., напр., [1]). Суть данной проблемы состоит в том, что энергия $E_{KS} = E(\text{LUMO}) - E(\text{HOMO})$ (LUMO, HUMO – низшая незанятая и верхняя занятая молекулярные орбитали) и определяемая как энергия запрещенной зоны Кона–Шэма существенно занижена по сравнению с экспериментальными значениями E_{exp} ($(E_{\text{exp}} - E_{KS}) / E_{\text{exp}} \sim 30$–$100$ %), что существенно затрудняет теоретический поиск новых материалов с заданными электронными (оптическими, фотовольтаическими, термоэлектрическими) свойствами. Ситуация осложняется отсутствием корреляции в отклонениях E_{KS} от E_{exp}. Различия в E_{KS} и E_{exp} при расчете в рамках локальных или квазилокальных приближений обменно-корреляционной энергии обычно связывают с разрывом функциональной производной обменно-корреляционной энергии [2] или ошибок нелокальной части обменно-корреляционной энергии [3]. Часто утверждают, что ширина запрещенной зоны определяется возбужденными состояниями и поэтому не может быть описана в рамках основного состояния, рассчитываемого DFT [4]. Много усилий было приложено для решения проблемы ширины запрещенной зоны путем введения дополнительных параметров [5], рассмотрения DFT, зависящего от времени (TDDFT) [6], точного обмена [7], гибридных и модифицированных потенциалов [8]. Наиболее последовательный (и точный) расчет ширины запрещенной зоны может быть выполнен с привлечением многочастичной теории возмущений (GW-метод) [9; 10]. Однако следует отметить, что для некоторых полупроводников ошибка определения ширины запрещенной зоны в рамках GW-метода составляет более 20 %, причем причины возникновения столь больших ошибок неясны. Кроме того, выполнение GW-расчетов требует привлечения очень больших вычислительных ресурсов и практически невыполнимо для структур с большим числом атомов.

В данном сообщении представлено решение проблемы расчета ширины запрещенной зоны в рамках метода DFT.

Вывод основных соотношений. Прежде всего отметим, что определение $E_{KS} = E(\text{LUMO}) - E(\text{HOMO})$ как ширины запрещенной зоны является недостаточно корректным. Действительно, как следует из основной теоремы Хоэнберга–Кона и уравнений Кона–Шэма энергии орбиталей и волновые функции Кона–Шэма (ε_i, φ_i) не имеют какого-либо физического смысла. Данные величины носят вспомогательный характер и служат для точного определения полной энергии системы как функции плотности электронов, и только энергия самой верхней заполненной орбитали E(HOMO), отсчитанная от вакуумного нуля, – это энергия ионизации [11]. Поэтому получим для ширины запрещенной зоны (E_g) выражение, используя только полную энергию системы как функционала плотности заряда $E[n(\vec{r})]$

$$E[n(\vec{r})] = T_s[n(\vec{r})] + E_{\text{ext}}[n(\vec{r})] + E_H[n(\vec{r})] + E_{xc}[n(\vec{r})] = T_s[n(\vec{r})] + \int v(\vec{r})n(\vec{r})d\vec{r} + \frac{1}{2}\int \frac{n(\vec{r})n(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}d\vec{r}' + E_{xc}[n(\vec{r})], \quad (1)$$

где $T_s[n(\vec{r})]$ – функционал кинетической энергии невзаимодействующих электронов; $E_{\text{ext}}[n(\vec{r})]$ – функционал энергии электронов во внешнем поле; $E_H[n(\vec{r})]$ – функционал энергии кулоновского взаимодействия электронов; $E_{xc}[n(\vec{r})]$ – функционал обменно-корреляционной энергии электронов. Для расчета ширины запрещенной зоны рассмотрим электронную конфигурацию системы, схематично представленную на рис. 1. Структура I представляет собой две кристаллические расширенные элементарные ячейки (КРЭЯ) (a) и (b) в зарядовом состоянии $Z = 0$, а в структуре II эти ячейки находятся в зарядовых состояниях $Z = +2$ и $Z = -2$ соответственно. Разложим функционал полной энергии системы на сумму функционалов энергии пары электронов, занимающих HOMO орбиталь и функционал энергии КРЭЯ в зарядовом состоянии $Z = +2$ ($E_0[n(\vec{r})]$). Тогда для полной энергии структуры I ($E_I[n(\vec{r})]$) и структуры II ($E_{II}[n(\vec{r})]$) получаем

$$E_I[n(\vec{r})] \equiv 2E[n(\vec{r}), Z = 0] = 2(E_0[n(\vec{r})] + E_{(1,0)}[n(\vec{r})] + E_{(\uparrow\downarrow,0)}[n(\vec{r})]); \quad (2)$$

$$E_{II}[n(\vec{r})] \equiv (E[n(\vec{r}), Z = +2] + E[n(\vec{r}), Z = -2]) = 2E_0[n(\vec{r})] + E_{(1,0)}[n(\vec{r})] + E_{(\uparrow\downarrow,0)}[n(\vec{r})] + E_{(12,E_g)}[n(\vec{r})] + E_{(2,E_g)}[n(\vec{r})] + E_{(\uparrow\downarrow,E_g)}[n(\vec{r})] + E_{(21,E_g)}[n(\vec{r})] + 2E_g, \quad (3)$$

где $E[n(\vec{r}), Z]$ – функционал энергии КРЭЯ в данном зарядовом состоянии Z; $E_0[n(\vec{r})]$ – функционал энергии, описывающий взаимодействие всех электронов (без пары электронов 1 или 2) друг с другом и ионами КРЭЯ; $E_{(1,0)}[n(\vec{r})]$, $E_{(2,E_g)}[n(\vec{r})]$ – функционалы энергии, описывающие взаимодействие пары электронов HOMO орбитали потолка валентной зоны и дна зоны проводимости со всеми электронами и ионами КРЭЯ; $E_{(\uparrow\downarrow,E_g)}[n(\vec{r})]$, $E_{(\uparrow\downarrow,0)}[n(\vec{r})]$ – функционалы энергии, описывающие взаимодействие электронов на данной HOMO орбитали; $E_{(12,E_g)}[n(\vec{r})]$, $E_{(21,E_g)}[n(\vec{r})]$ – функционалы энергии, описывающие изменение энергии 1(2) пары электронов при взаимодействии со 2(1) парой; E_g – ширина запрещенной зоны. Для исключения из расчетов большой энергии $E_0[n(\vec{r})]$ будем анализировать разность $E_{II}[n(\vec{r})] - E_I[n(\vec{r})]$. Предположим, нам известен точный функционал $E_{xc}^{(0)}[n(\vec{r})]$ и функционал $E_{xc}^{(\text{LDA})}[n(\vec{r})]$ локального приближения (LDA) для обменно-корреляционной энергии. Рассмотрим выражение

$$\Delta_{xc}^{\infty} = (E_{II}[n(\vec{r})] - E_I[n(\vec{r})])^{(0)} - (E_{II}[n(\vec{r})] - E_I[n(\vec{r})])^{(\text{LDA})}. \quad (4)$$

С учетом (2), (3) Δ_{xc}^{∞} имеет вид

$$\Delta_{xc}^{\infty} = 2E_g^{(0)} - 2E_g^{(\text{LDA})} + \left(E_{(2,E_g)}^{(0)}[n(\vec{r})] - E_{(2,E_g)}^{(\text{LDA})}[n(\vec{r})]\right) - \left(E_{(1,0)}^{(0)}[n(\vec{r})] - E_{(1,0)}^{(\text{LDA})}[n(\vec{r})]\right) + \left(E_{(\uparrow\downarrow,E_g)}^{(0)}[n(\vec{r})] - E_{(\uparrow\downarrow,E_g)}^{(\text{LDA})}[n(\vec{r})]\right) - \left(E_{(\uparrow\downarrow,0)}^{(0)}[n(\vec{r})] - E_{(\uparrow\downarrow,0)}^{(\text{LDA})}[n(\vec{r})]\right) + 2\left(E_{(21,E_g)}^{(0)}[n(\vec{r})] - E_{(21,E_g)}^{(\text{LDA})}[n(\vec{r})]\right). \quad (5)$$

Как следует из (1), разность функционалов вида $E^{(0)}[n(\vec{r})] - E^{(\text{LDA})}[n(\vec{r})]$ определяется обменно-корреляционной энергией $E^{(0)}[n(\vec{r})] - E^{(\text{LDA})}[n(\vec{r})] = E^{(0,xc)}[n(\vec{r})] - E^{(\text{LDA},xc)}[n(\vec{r})]$. Разложим обменно-корреляционную энергию на локальную и нелокальную части $E^{(0,xc)}[n(\vec{r})] = E^{(\text{LDA},xc)}[n(\vec{r})] + E^{(\infty,xc)}[n(\vec{r})]$ и представим (5) в виде

$$\Delta_{xc}^{\infty} = 2E_g^{(0)} - 2E_g^{(\text{LDA})} + 2E_{(21,E_g)}^{(\infty,xc)}[n(\vec{r})] + \left(E_{(2,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]\right) + \left(E_{(\uparrow\downarrow,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(\uparrow\downarrow,0)}^{(\infty,xc)}[n(\vec{r})]\right). \quad (6)$$

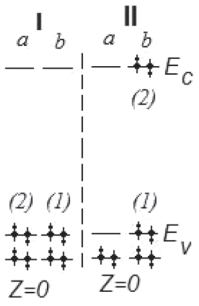

Рис. 1. Электронная структура модельной системы. I – две ячейки КРЭЯ (a) и (b) в зарядовом состоянии Z = 0; II – ячейки КРЭЯ в зарядовых состояниях a: $Z = +2$ и b: $Z = -2$; (1) и (2) – обозначения пар электронов на HOMO орбитали, соответствующие формулам текста

Из (6) получаем точное соотношение для ширины запрещенной зоны в рамках метода DFT, выраженное через ширину запрещенной зоны, рассчитанную в локальном (LDA) приближении обменно-корреляционной энергии:

$$E_g^{(0)} = E_g^{(\mathrm{LDA})} + \frac{\Delta_{xc}^{\infty}}{2} - \frac{1}{2}\left\{\begin{array}{l} 2E_{(21,E_g)}^{(\infty,xc)}[n(\vec{r})] + \left(E_{(2,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]\right) + \\ + \left(E_{(\uparrow\downarrow,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(\uparrow\downarrow,0)}^{(\infty,xc)}[n(\vec{r})]\right) \end{array}\right\}. \quad (7)$$

Из (7) сразу же следует, что LDA приближение не приводит к правильным значениям ширины запрещенной зоны, поскольку ширина запрещенной зоны определяется также нелокальной составляющей обменно-корреляционной энергии. Как уже было отмечено выше, для выполнения расчетов на основании (7) мы должны выразить все слагаемые через функционалы вида $E[n(\vec{r}), Z]$. В нашем случае ширина запрещенной зоны в приближении LDA имеет вид

$$E_g^{(\mathrm{LDA})} = \frac{1}{2}\left\{ 2E^{(\mathrm{LDA})}[n(\vec{r}), Z=0] - (E^{(\mathrm{LDA})}[n(\vec{r}), Z=+2] + E^{(\mathrm{LDA})}[n(\vec{r}), Z=-2])\right\}. \quad (8)$$

Величина Δ_{xc}^{∞} также может быть рассчитана через функционалы вида $E[n(\vec{r}), Z]$ (см. (4) и дальнейшее обсуждение). Последнее слагаемое в (7)

$$F^{(\infty,xc)}[n(\vec{r}), E_g] = \frac{1}{2}\left\{\begin{array}{l} 2E_{(21,E_g)}^{(\infty,xc)}[n(\vec{r})] + \left(E_{(2,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]\right) + \\ + \left(E_{(\uparrow\downarrow,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(\uparrow\downarrow,0)}^{(\infty,xc)}[n(\vec{r})]\right) \end{array}\right\} \quad (9)$$

выражено через нелокальную часть функционала обменно-корреляционной энергии, и точный аналитический вид данного слагаемого получить нельзя [11]. Тем не менее, для $F^{(\infty,xc)}[n(\vec{r}), E_g]$ можно получить аппроксимирующее выражение на основании асимптотического поведения. Прежде всего, отметим, что в дальнейшем для нелокального приближения обменно-корреляционной энергии мы будем полагать выполненным условие $(E_{(\uparrow\downarrow,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(\uparrow\downarrow,0)}^{(\infty,xc)}[n(\vec{r})]) = O(E_{(2,E_g)}^{(\infty,xc)}[n(\vec{r})] - E_{(1,0)}^{(\infty,xc)}[n(\vec{r})])$, которое с очевидностью выполняется для малых значений ширины запрещенной зоны.

В случае $E_g << 1$ (полагаем, что E_g нормировано) имеем

$$\lim_{E_g \to 0}(F^{(\infty,xc)}[n(\vec{r}), E_g]) \cong E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]. \quad (10)$$

Представим $F^{(\infty,xc)}[n(\vec{r}), E_g]$ в виде $F^{(\infty,xc)}[n(\vec{r}), E_g] = E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]\alpha(E_g)$, где $\alpha(E_g)$ – неизвестная безразмерная функция ширины запрещенной зоны. Для больших значений ширины запрещенной зоны $E_g >> 1$ функция $\alpha(E_g)$ должна быть медленно меняющейся функцией E_g и стремиться к асимптотическому значению $\alpha(E_g >> 1) \cong \mathrm{const} \equiv \alpha_0$. Легко видеть, что в этом случае зависимость $F^{(\infty,xc)}[n(\vec{r}), E_g]$ как функцию ширины запрещенной зоны можно представить в виде

$$F^{(\infty,xc)}[n(\vec{r}), E_g] = E_{(1,0)}^{(\infty,xc)}[n(\vec{r})][\alpha_0 + e^{-f(E_g)}], \quad (11)$$

где $f(E_g) > 0$ и монотонно возрастает с ростом ширины запрещенной зоны. Разлагая в ряд Тейлора и ограничиваясь линейным приближением мы получаем $f(E_g) = f(0) + (\partial_{E_g} f)_{E_g=0} E_g$. Производная $(\partial_{E_g} f)_{E_g=0} \equiv 1/E_0$ является параметром метода и может быть выбрана на основании сопоставления с экспериментом. Однако при проведении конкретных расчетов параметр E_0 не оптимизировался, а полагался равным $E_0 = 1$. Величина α_0, определенная на основании асимптотики (10), равна $\alpha_0 = 1$. Окончательно для ширины запрещенной зоны получаем

$$E_g^{(0)} = E_g^{(\mathrm{LDA})} + \frac{\Delta_{xc}^{\infty}}{2} - \frac{E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]}{2}\left(1 + e^{-\frac{E_g}{E_0}}\right). \quad (12)$$

Выражение для функционала $E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]$ через функционалы вида $E[n(\vec{r}), Z]$ может быть получено на основании (1), (2):

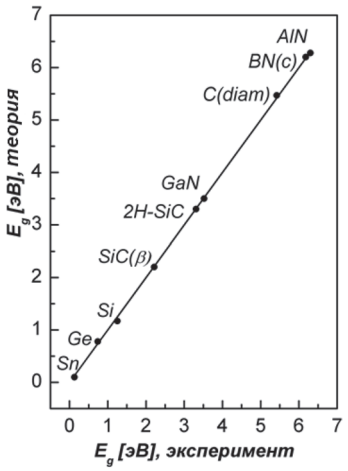

Рис. 2. Рассчитанные значения ширины запрещенной зоны как функция экспериментальных значений для исследованного ряда полупроводников

$$E_{(1,0)}^{(\infty,xc)}[n(\vec{r})] = \left\{E^{(0)}[n(\vec{r}), Z=0] - E^{(\mathrm{LDA})}[n(\vec{r}), Z=0]\right\} - \left\{E^{(0)}[n(\vec{r}), Z=+2] - E^{(\mathrm{LDA})}[n(\vec{r}), Z=+2]\right\}. \quad (13)$$

Как следует из (4) и (13), для вычисления ширины запрещенной зоны необходимо проводить вычисление функционалов с точным значением обменно-корреляционной энергии $E_{xc}^{(0)}[n(\vec{r})]$, которое неизвестно [11]. При проведении конкретных расчетов в качестве приближения для $E_{xc}^{(0)}[n(\vec{r})]$ выбирался широко используемый гибридный функционал B3LYP [12]. Для данного функционала выполняется условие разложения на локальную и нелокальную части [12]

$$E_{xc}^{(0)}[n(\vec{r})] \cong E_{xc}^{(\mathrm{B3LYP})}[n(\vec{r})] = E_{(x)}^{(\mathrm{LDA})} + a_0(E_{(x)}^{(\mathrm{HF})} - E_{(x)}^{(\mathrm{LDA})}) + a_x(E_{(x)}^{(\mathrm{GGA})} - E_{(x)}^{(\mathrm{LDA})}) + E_{(c)}^{(\mathrm{LDA})} + a_c(E_{(c)}^{(\mathrm{GGA})} - E_{(c)}^{(\mathrm{LDA})}). \quad (14)$$

Расчет E_g для ряда одноатомных и двухатомных полупроводников. На основании (12) нами был выполнен расчет ширины запрещенной зоны для следующего ряда полупроводников: Sn(Fd3m), Ge(Fd3m), Si(Fd3m), C(Fd3m), BN(c)(F43m), SiC(β)(F43m), 2H-SiC($P6_3mc$), AlN($P6_3mc$), GaN($P6_3mc$) (в скобках указана симметрия). Ширина запрещенной зоны в LDA приближении (8) рассчитывалась в модели кристаллической расширенной элементарной ячейки. Размер КРЭЯ составлял 216 атомов с симметрией Fd3m, F43m и 256 атомов с симметрией $P6_3mc$. Псевдопотенциал выбирался в форме Si(C, Ge, N, B, Al).pz-vbc.UPF, Sn.pz-bhs.UPF. Геометрия КРЭЯ оптимизировалась путем минимизации полной энергии по координатам всех атомов. Суммирование проводилось по 14 точкам зоны Брюллюэна (сетке 3 × 3 × 3). Кинетическая энергия обрезалась при 272 эВ. Равновесная конфигурация атомов достигалась, когда величина силы на любом из атомов структуры становилась менее $2 \cdot 10^{-3}$ эВ/Å. Расчет функционалов Δ_{xc}^{∞} и $E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]$ был выполнен в кластерном приближении. Кластер состоял из ~80 атомов и строился на основании использованных в LDA расчетах расширенных кристаллических ячеек. Оборванные связи на границах кластера насыщались атомами водорода. Волновая функция кластера представлялась в базисе STO 6-31G. Структура кластера оптимизировалась.

На рис. 2 представлено сопоставление рассчитанных и экспериментальных значений ширины запрещенной зоны для исследованного ряда полупроводников. Видно, что для достаточного большого интервала значений E_g (~0–6 эВ) рассчитанные значения ширины запрещенной зоны практически совпадают с экспериментальными значениями как для прямозонных (AlN), так и непрямозонных изоструктурных (Sn, Ge, Si, C) полупроводников, а также для полупроводников с одинаковым составом, но разной структурой (SiC(β)(F43m), 2H-SiC($P6_3mc$)). Отметим, что величина поправок Δ_{xc}^{∞} и $E_{(1,0)}^{(\infty,xc)}$ существенно влияет на величину запрещенной зоны. Так, в таблице представлены рассчитанные поправки к $E_g^{(\mathrm{LDA})}$ для изоструктурного ряда полупроводников Ge, Si, C.

Рассчитанные значения ширины запрещенной зоны и поправок нелокальной составляющей обменно-корреляционной энергии для изоструктурного ряда полупроводников Ge, Si, C

Полупроводник	Ge	Si	C
$E_g^{(\mathrm{LDA})}$, [эВ]	0,89	0,87	4,5
Δ_{xc}^{∞}, [эВ]	0,20	0,65	1,11
$E_{(1,0)}^{(\infty,xc)}$, [эВ]	0,43	0,38	0,43
$\frac{\Delta_{xc}^{\infty}}{2} - \frac{E_{(1,0)}^{(\infty,xc)}[n(\vec{r})]\left(1+e^{-\frac{E_g}{E_0}}\right)}{2}$	–0,11	0,41	0,89
E_g, [эВ]	0,78	1,28	5,39

Предложенный метод дает возможность анализа энергии ионизации дефектов, вносящих глубокие уровни в запрещенную зону полупроводников (локализованных состояний) и электронных свойств наноструктур. Например, для хорошо известного дефекта в кремнии вакансия–кислород (А-центр) рассчитанная нами энергия ионизации дефекта составила E_c = 0,174 эВ (T = 0 К), что хорошо согласуется с экспериментальными значениями $E_c \sim 0{,}17$ эВ.

Заключение. В рамках теории функционала плотности развит метод расчета ширины запрещенной зоны полупроводников. Выполненный расчет ширины запрещенной зоны для ряда одноатомных и двухатомных полупроводников показал, что метод позволяет получить значения ширины запрещенной зоны практически с экспериментальной степенью точности. Важным является тот факт, что развитый метод может быть использован также для расчета как локализованных состояний (энергии глубоких уровней ионизации дефектов в кристаллах), так и электронных свойств наноструктур.

Исследование выполнено в рамках программы «Функциональные и композиционные материалы, наноматериалы».

Автор благодарит суперкомпьютерный центр ОИПИ НАН Беларуси.

Литература

1. *Xiao Zheng* et al. // Phys. Rev. Lett. 2011. Vol. 107, N 2. P. 026403-1–026403-4.
2. *Sham L. J., Schlüter M.* // Phys. Rev. Lett. 1983. Vol. 51, N 20. P. 1888–1890.
3. *Paula Mori-Sánchez, Aron J. Cohen, Weitao Yang* // Phys. Rev. Lett. 2008. Vol. 100, N 14. P. 146401-1–146401-4.
4. *Godby R. W., Schlüter M., Sham L. J.* // Phys. Rev. B 1987. Vol. 35, N 8. P. 4170–4171(R).
5. *Chan M. K. Y., Ceder G.* // Phys. Rev. Lett. 2010. Vol. 105. P. 196403-1–196403-3.
6. *Runge E., Gross E. K. U.* // Phys. Rev. Lett. 1984. Vol. 52, N 12. P. 997–1000.
7. *Stadele M.* et al. // Phys. Rev. B 1999. Vol. 59, N 15. P. 10031–10042.
8. *Tran F., Blaha P.* // Phys. Rev. Lett. 2009. Vol. 102, N 22. P. 226401-1–226401-4.
9. *Hedin L.* // Phys. Rev. 1965. Vol. 139, N 3A. P. 796–823.
10. *Camargo-Martinez J. A., Baquero R.* // Phys. Rev. B. 2012. Vol. 86. P. 195106.
11. *Кон В.* // УФН. 2002. Т. 172, № 3. С. 336–348.
12. *Kim K., Jordan K. D. J.* // Phys. Chem. 1994. Vol. 98, N 40. P. 10089–10094.

V. E. GUSAKOV

gusakov@ifttp.bas-net.by

CALCULATION OF THE BAND GAP OF SEMICONDUCTORS WITHIN THE FRAMEWORK OF THE DENSITY FUNCTIONAL METHOD

Summary

Within the framework of the density functional theory, the method was developed to calculate the band gap of semiconductors. Calculation of the band gap for a number of monoatomic and diatomic semiconductors demonstrated that the method gives the value of the band gap of almost experimental accuracy. An important point is the fact that the developed method can also be used to calculate both localized states (energy deep-level of defects in crystal), and electronic properties of nanostructures.